Large-Scale Content Analysis of Historical Newspapers in the Town of Gorizia 1873-1914

Nello Cristianini, Thomas Lansdall-Welfare, Gaetano Dato

University of Bristol, Bristol, United Kingdom

Emails:

nello.cristianini@bristol.ac.uk

thomas.lansdall-welfare@bristol.ac.uk

g.dato@bristol.ac.uk

Corresponding author:

Nello Cristianini <nello.cristianini@bristol.ac.uk>

Merchant Venturers Building, University of Bristol, Woodland Road, Bristol BS8 1UB, United Kingdom

Acknowledgments:

Two of the authors (NC, TLW) have been supported by the ThinkBIG ERC project. The authors thank Dr. Marco Menato, Director of the Biblioteca Civica Isontina, for his invaluable help, support and encouragement. We also thank the Library itself for making the microfilms available; the FindMyPast digitisation team for their support and advice during the digitisation process; the Slovenian digital library, dlib.si, for making their data available online for academic research; and Dr. Gianluca Demartini - of the University of Sheffield - for help and discussions about information extraction, that allowed us to inform this study and plan future ones. The authors thank Igor Devetak and Andrea Sgarro for their advice and help reading an early draft.

Abstract. We have digitised a corpus of Italian newspapers published in 1873-1914 in Gorizia, the county town of an area in the North Adriatic at the crossroad of the Latin, Slavic and Germanic civilizations, then part of the Habsburg Empire and now divided between Italy and Slovenia. This new corpus (of 47,466 pages) is analysed along with a comparable set of local Slovenian newspapers, already digitised by the Slovenian National Library. This large and multilingual effort in digital humanities reveals the statistical traces of events and ideas that shaped a remarkable place and period. The emerging picture is one of rapid cultural, social and technological transformation, and of rising national awareness, combining the larger European pattern with uniquely local aspects.

Keywords: Austro-Hungarian Empire, Digital Humanities, Digital Newspaper Archives, Gorizia

1 - Gorizia, Gorica, Görz

Gorizia lays in the North Adriatic region, an area where Latin, Slavic and Germanic civilizations met, traded and battled for centuries. The town and its territory are inhabited by Italian and Slovenian speakers, and from 1500 to the First World War (WW1) were part of the Habsburg Empire. For centuries, it was on the western border of the Habsburg monarchy, while during the 19th c. and after each of the World Wars the border moved, leaving the region split between Italy and Slovenia, with the modern-day town remaining mostly in Italy.

Founded in the high middle ages[1], for most of its history Gorizia has been a border territory, with an ethnic make-up that varied according to location and time. As the borders moved significantly through the ages, both as the result of wars and as the result of internal re-organisations, we will focus on the territory of Gorizia as defined by its main landmark: the course of the river known as the *Isonzo* in Italian and *Soča* in Slovenian, which crosses both the city and its territory and runs for 136 km from the Julian Alps to the Adriatic Sea. It flows from the most sacred mountain of the Slovenian people, the Triglav, 50 km north of Gorizia, to the Adriatic Sea circa 25 km south of the

Figure 1 - A map of the North Adriatic region in the period between 1866 and 1918. The "Princely County of Gorizia and Gradisca" is highlighted in green, the other regions of the Austrian-Hungarian Empire in pale green, and the Kingdom of Italy in pale orange. The blue line is the river *Isonzo* (*Soča*), and at its centre is the town of Gorizia (*Gorica*, *Görz*).

town, in a historically Venetian area. Thus, Gorizia represents a point of ethnic transition along the course of the river itself.

During the period under investigation (1873-1914), the entire course of the river and its surrounding territory were part of the Princely County of Gorizia and Gradisca[2], a crown land within Austria-Hungary, one of the three regions forming the Austrian Littoral[3], an administrative entity first created in 1849 that also included Trieste and the Istrian Peninsula.

Since the late middle ages, the urban centre of Gorizia was mostly inhabited by an Italian population, with the surrounding countryside mostly inhabited by Slovenians who call the city *Gorica*. Until WW1, there was also a small population of Austrians, who called the city *Görz*. The Italian component was further linguistically divided between Venetian and Friulian speakers, who were mostly situated to the south near the coast and over the plains west of the city respectively, with some Italians speaking a combination of the two. It is worthwhile to remark that the land roughly between the river *Isonzo* and the river *Tagliamento* to its west is called Friuli. This name, together with *Isonzo-Soča* and the translations for Littoral in the languages of the region, often features in local newspapers' names.

This urban Italian population were often members of the middle class or the aristocracy, while the surrounding countryside was predominantly Slovenian farmers who spoke a local Slovenian dialect. The Austrian population in the city included many functionaries for the central government, along with a contingent of entrepreneurs and business people. In the 1800s, especially after the Habsburg empire's loss of the western Italian territories in 1866, tourists from the colder German speaking regions of the empire started spending long periods - and even retiring - in the town because of Gorizia's fame for having a good climate[4].

In the late 19th c., various developments led to a change in the old ethnic-social equilibrium, including changes in electoral laws that gave increased representation to lower income households, increased tolerance of the expressions of the cultural life of Italians and Slovenes by the authorities, and increasing urbanization of the countryside's population. This resulted in national identities being pushed into a higher level of conflict, with inter-ethnic cohesion of the multinational Habsburg empire moving from crisis to crisis towards WW1, in Gorizia as much as in the rest of the country. In Gorizia however, national conflict was never particularly violent in the period under investigation (Fabi 1991, 9-10, 32-35; Ferrari 2002, 313-8; Marušič 2005, 7-12).

1.1 - Times of change

The decades that straddle the 19th and 20th century were times of great social and technological change, not only in Gorizia, but in the entire world, coinciding with a particular stage in the development of the modern world: the second industrial revolution and the dawn of mass society. For Gorizia, those years were even more crucial, as they led up to WW1 and the end of centuries of Habsburg rule.

Census data signals a continuous increase in the urban population of Gorizia, following a general trend in the western world of massive migration

from the countryside to cities, which for Gorizia also had the effect of changing the ethnic make-up too. The number of registered residents grew from 16,659 in 1869 to 29,291 in 1910, when the last census before WW1 was taken. Nonetheless, the increase accelerated in the last decade, with the population still at 23,765 individuals in 1900. Italian speakers decreased from 16,112 in 1900, to 14,812 in 1910, while Slovenian speakers surged from 4,754 to 10,790 - if we trust the methods adopted by the Habsburg census bureau for language speaking recognition. While in the town the Italians were in a decreasing majority, Slovenian speakers were always the majority in the rest of the County, where in the last Habsburg census there were 154,564 Slovenian speakers versus 90,146 Italian speakers. Other factors were also at play, including migration being directed overseas rather than into the cities, while falling mortality rates increased the pressure of urbanization (Fabi 1991, 252-4; Kalc 2013, 684-701; Marušič 2005, 45-46).

These were years characterised by fast technological development, enhanced communications, the emergence of ideologies, national identities, and a general faith in social progress that coexisted with pervasive anxiety and fear about these same transformations. One of the areas that blossomed in the period under investigation was the Press, due mostly to reforms in the laws regulating freedom of speech and association, as will be described below.

While we analyse the contents of the newspapers of that period, we will pay particular attention to the signs of cultural and social change: the arrival of new technologies, ideas, opportunities and problems into this part of Europe. We will also focus on how these signs of cultural and social change interacted with the aspects that are specific to this region. We are interested in how a large-scale corpus analysis of press content - both based on statistical trends and on close reading - can help us understand the change that affected this crucial part of Europe in the years leading up to WW1.

1.2 – Political Landscape and the Press

The political landscape in the North Adriatic border region in the 19th and 20th c. was influenced by two main features: the ideological and the national. These two features are key to understanding the local politics in the years up until the end of the Cold War. There were clear, distinctive regional and age sub-patterns to be understood in this large geospatial timeframe.

In the city of Gorizia in the 1870s, the main ideologies were liberalism and Catholic conservatism. The Rerum Novarum encyclical issued by Pope Leo XIII in 1891 led to most Catholics embracing the Christian Social thought throughout the 1890s. This was a response by the Church to the expansion of secularism and socialism that advocated for a deeper political involvement of the Catholics in defence of the lower classes. The Christian Social movement was quite successful in the county of Gorizia; socialists had only a major stronghold at the shipyards of Monfalcone in the early 1900s.

National identities followed the fracture among the main ethnic groups of Gorizia. Italian and Slovenian were the two dominant national identities, with a small representation of the Austrian-German identity. Being structured along the ideological and the national lines, there were at

least four political orientations in Gorizia, satisfying all the possible political combinations, plus the tiny group of Austrian local politicians.

National identities also influenced another feature that was meaningful in the timeframe of this analysis: the support professed to the central government and to a united Austria-Hungary. As we move towards WW1, non-German nationalities sought deeper forms of autonomy and even separatism, as was done by some of the more radical Italian groups called the "*irredentisti*" (Fabi 1991, 12-46; Ferrari 2002, 340-75; Kacin-Wohinz and Troha 2000[5], 69-79; Marušič 2005, 239-344).

The press, forced to remain within certain boundaries by the authorities, gave voice to the different stakeholders in the politics of Gorizia, each of which aspired to present the public with its own perspective. The newspapers that ended up lasting the longest catered for one or more of the four quadrants defined by the two prevailing political dimensions of the time: liberal vs Catholic; Italian vs. Slovenian.

La Gazzetta Goriziana was the first periodical printed in Gorizia, a weekly sold between 1774 and 1775. More local journals became available in the following years, but they were mostly short lived. A German press existed in certain periods, but it relied on a smaller readership and it could not sustain itself in the long run. However, as education improved and social complexity increased, by the second half of the 19th c. both Italian and Slovenian communities had daily newspapers, a regular audience and their own print shops (De Grassi 1982, 55; Feresin 2007, 14-17; Gorian 2010). The process was boosted by the constitutional reforms of December 1867, when laws governing the freedom of speech were changed, enabling the birth of multiple newspapers that represented the various political and ethnic positions during the last quarter of the century (Ferrari 2002, 342; Horel 2015, 88-90).

Here we describe the five longest lasting newspapers of the town, covering most of the period from 1873 to 1914. They include the four main political positions, as shown in Table 1, where the political position of each periodical is indicated, while the lifespans of each publication are shown in Table 2.

Table 1 - The political position of the five newspapers under study.

	Liberal	Catholic
Italian	Il Corriere Friulano / Corriere di Gorizia	L'Eco del Litorale
Slovenian	Soča (1871-1899) Soča (1899-1914)	Soča (1871-1899) Gorica Primorski List

Table 2 - Number of different issues, words and years available for each corpus.

Corpus	**Words**	**Days / Issues**	**Years**
Gorica	16,189,845	1472	16 (1899-1914)
Primorski list	9,662,362	892	22 (1893-1913)
Soča	38,153,317	3342	44 (1871-1914)
EDL	50,429,470	6356	42 (1873-1915)
CFG	67,068,101	5475	31 (1883-1914)

Corriere di Gorizia / Il Corriere Friulano (CFG). While the Italian liberals were in control of the municipality in the period under investigation, they suffered some opposition from the central authorities about regularly printing their own newspapers, due to their political line over the relationship with Rome, Vienna and the Slovenes. *L'Isonzo* was the first main liberal Italian newspaper after the 1867 regulations of the Austro-Hungarian empire. It existed from October 1871 to 1880, when authorities forcefully closed it. Later, many of its stakeholders, headed by the Italian Jewish intellectual Carolina Luzzatto, founded the daily *Il Corriere di Gorizia* in its place in 1883. By 1899, authorities compelled it to change into *Il Friuli Orientale* for one year. The group led by Luzzatto also began printing *Il Corriere Friulano* in 1901, and after a few months of coexistence with *Il Friuli Orientale*, the latter was shut down and *Il Corriere Friulano* continued as the publication representing Italian liberals' views up until 1914, when it was finally closed following the enforcements on press restrictions due the state of war (De Grassi 1982, 58-60, 70-1; Ferrari 2002, 356-7, 366).

Our corpus includes the whole series of *Corriere di Gorizia* and *Il Corriere Friulano* but not *L'Isonzo* nor *Il Friuli Orientale*, thus covering the period from 1883 to 1914, with a gap of 16 months from January 1900 to April 1901. These two missing publications are available as part of the microfilm collection of the Library of Gorizia, and could be added to the corpus in the future.

L'Eco del Litorale (EDL). In October 1871, two weeks after the birth of the Italian liberal publication *L'Isonzo*, the diocese of Gorizia supported the birth of the Italian newspaper *Il Goriziano* (not to be confused with a later, radical Italian liberal magazine of 1876-78). It became *L'Eco del Litorale* from 1873 onward, with locally well-known father Domenico Alpi being one of its key writers for most of its years. It was initially an anti-liberal supporter of the empire and the clergy; its stronghold was the countryside. In the 1890s, it gradually embraced the Christian Social turn after the publication of the Rerum Novarum encyclical, and became supportive of father Luigi Faidutti, who led the Christian Social movement. *L'Eco del Litorale* was the main target of Luzzato's newspapers polemics, but was also the target of state control and the occasional temporary suspensions[6]. The newspaper was published daily, bi-weekly and tri-weekly at various times, and in April 1915, it moved to Vienna then Trieste, finally ending publication in 1918 (De Grassi 1982, 74-5; Feresin 2007, 17; Medeot 1981, 29-40).

We digitised the whole series of *L'Eco del Litorale* from 1874 to spring 1915 (the part edited in Gorizia), but we have not included *Il Goriziano*, which is available as microfilm at the Library of Gorizia and could be added to the corpus (De Simone 1996). Figure 2 shows an example front page from each of the two Italian newspapers that we digitised as part of this study.

Soča, Gorica, Primorski List. Liberal and Catholic Slovenes forged an uneasy alliance under the Slovenian colours, publishing the influential newspaper, the *Soča*, along with creating an association called *Sloga* from 1875. The newspaper was born in March 1871 and tried, with no lack of hardships, to represent a multifaceted Slovenian


IL CORRIERE FRIULANO
GIORNALE POLITICO-LETTERARIO QUOTIDIANO
L'ECO DEL LITORALE

Figure 2 – Examples of front pages for *Il Corriere Friulano* and *L'Eco del Litorale*.

political environment, split by generational and ideological divides. These also affected the Catholic area, fragmented into the clerical-conservative and the progressive groups that later in the 1890s fully embraced the Christian Social turn. However, apart from the splits in 1872-75 and 1889-92, when the Catholic component tried to run their own newspapers, the liberal-Catholic alliance survived until 1899 when the fracture was officially confirmed. Christian-Social Slovenes started publishing the *Gorica* in the same year, led by the key political figure of Anton Gregorčič, *Soča*'s director in 1882-89 and 1892-99. From 1899, *Soča* remained in the hands of the liberal leaders Andrej Gabrščček and Henrik Tuma (Tuma later turned socialist and abandoned *Soča* in 1908) until it concluded its publication in 1915.

Conservative Catholics founded the *Primorski List* in 1893 in Trieste to create a magazine devoted to all the Catholics in the Littoral (Sl: *Primorje*), while attacking the Christian Socialists as crypto-socialists. From 1894, *Primorski List* was printed in Gorizia, with a temporary alliance being forged with *Soča* in 1898, where both began supporting the Christian Social Slovenes. After the Liberal-Catholic break of 1899 and an attempted merger with *Gorica* in 1900, *Primorski List* remained a supporter of the Christian

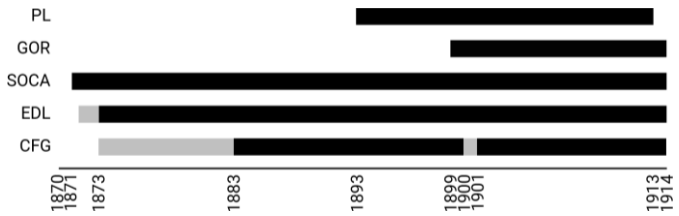

Figure 3 - Time coverage of the five newspapers included in this study. Note that we combine under CFG both *Corriere di Gorizia* and *Il Corriere Friulano* but not *L'Isonzo*. The gap in CFG correspond to *Il Friuli Orientale* which could be considered different expressions of the same publishing group. Note also that the grey gap in EDL at the beginning corresponds to the time when it was named *Il Goriziano*. The grey gaps represent newspapers that could be added to this collection.

Social thought. Missia, the archbishop of Gorizia, could not tolerate that both main Catholic journals of the city were supporting the Christian Social cause and the excessive divisions among Catholics. He therefore forced the closure of both *Primorski List* and *Gorica* at the beginning of 1914, replacing them with the *Goriški list*, which ended publication in 1915 because of the war, and which is not included in our research (Marušič 2005, 250-6, 330-44; Medeot 1981, 29-30).

These three newspapers, *Soča*, *Gorica*, and *Primorski List*, were all digitised by the National and University Library of Slovenia, and are available online at the Digital Library of Slovenia website (dLib.si 2017). Figure 3 shows the lifespan of all five newspapers included in this analysis.

2 - The Digital Corpus

2.1 - Italian Newspapers

Papers and Microfilms. The collections of periodicals were compiled by the *Biblioteca Statale Isontina* of Gorizia as they were published, and kept in the same building until today, surviving the bombings of two world wars. Between 1984 and 1999 the library embarked on a program to photograph these collections and preserve them on microfilm. This resulted in 563 microfilm reels, containing the periodicals most representative of the geographical and cultural area (excluding national newspapers, which were already being kept in other libraries). This work included filling whatever gaps were found in the collections by resorting to other local libraries, where possible, so that the microfilm collection is more complete than the paper collection, resulting in the creation of a valuable resource for future digitisation (De Simone 1996).

Image Digitisation. We digitized the whole series of *Il Corriere di Gorizia*, and *Il Corriere Friulano*, thus covering the period from 1883 to 1914, with a gap in 1900 (when it was called *Il Friuli Orientale*), and without *L'Isonzo*, from 1871 to 1900. In total, this produced digital images for 21,855 pages. We also digitised the whole series of *L'Eco del Litorale* from 1874 to 1914, producing digital images for 25,611 pages in total. We did not digitise its precursor *Il Goriziano*.

The images of the newspapers were stored on 42 reels of microfilm (100ft rolls of 35mm film, each capable of containing approximately 600 images of double pages), which were digitised using Wicks&Wilson 8850 microfilm scanners. The images produced were raw uncompressed grayscale

tiff images at 300 dpi, which were converted to jpegs for Quality Control (QC). The original material had been captured with large white borders around them, so these were cropped using an automated edge-detection process. Double-page images were split into single-page images using a manual process, carried out in batches just before QC.

QC processes included logical checks: matching the number of images on a reel to that of images in the final output folder and checking the file sizes. Physical checks of the images included visually checking for brightness and that images were cropped or cut correctly, as well as verifying that folder names and metadata are correct. Processing the images in a production workflow is key to quality delivery and was fully documented, to allow traceability of any issues or trends.

This phase produced a total of 47,466 images of individual newspaper pages (EDL - 25,611 pages, CFG - 21,855 pages).

Annotation of Images. Each of the 42 reels had approximately 1,130 single-page images, with the time interval covered by the newspapers varying from reel to reel (each reel had been filled as much as possible for reasons of economy). There was not an obvious automated procedure to apply the correct date to each of the pages (date is typically indicated only on the front page of a newspaper issue, and is not always readable; the number of pages per issue can vary, and the frequency of publication - days of the week - also changed over time). For these reasons, we resorted to hand-labelling each of the 47,466 pages with their date, which was a very time-consuming procedure. Errors are possible as this phase of the processing was done by hand, without a second phase of Quality Control. However, we did perform basic consistency checks and we are reasonably confident in this hand labelling of dates to the newspaper pages: the order of the dates is consistent with the physical ordering of the images in the reels; there are no long gaps without any pages; no dates are associated with an abnormal number of pages; the day of the week matches the date. This level of annotation is not perfect but fits the purposes of an analysis of statistical trends (we work at the level of trimesters, as described in later sections).

This annotation enabled us to establish that the CFG part of the corpus contains 5,475 distinct issues (distinct dates) and the EDL section of the corpus contains 6,356 issues (distinct dates).

Optical Character Recognition (OCR). We extracted digital text from the images using Abbyy FineReader version 12[7], while specifying that the text was in Italian. FineReader attempts (where possible) to respect the boundaries between columns of text on a page, and therefore to segment the page into articles, but we considered each page as a single unit of analysis, without going down to the level of individual articles, nor did we attempt to extract titles. Also, while FineReader does attempt to distinguish between articles, tables and images, we used all text (tables and articles), and did not make any use of images detected by the OCR tool. Future work could considerably improve on this process by trying to identify individual articles, or individual tables. This step enabled us to estimate that the CFG part of the corpus contains 67,068,101 words and the EDL part contains 50,429,470 words.

As our version of FineReader does not report estimates of recognition accuracy, we directly estimated the error rate of the OCR process by randomly selecting 10 articles, representing both outlets and different decades, before transcribing them by hand to obtain a ground truth to compare with the output from FineReader. Of the 10 randomly selected articles, one was not processed by FineReader as it contained a large image, leading to the entire article being treated as an image and so was not passed through the OCR process by the software. We did not consider this article in our estimation of error rates. In the remaining 9 articles, the Character Error Rate (CER) was found to be 24.6% and the Word Error Rate (WER) 23.1% (which is consistent with similar estimates in different projects, *e.g.* (Lansdall-Welfare et al. 2017)). These errors can be due to preservation issues (*e.g.* dirt on pages or on the film), scanning issues (*e.g.* ink visible from the other side of the page), or just confusion due to print ambiguities (*e.g.* the letters 'e' and 'o' are similar looking).

We decided not to use sophisticated techniques from Natural Language Processing (NLP) or Information Extraction (IE) because they would need either cleaner text or a larger corpus. NLP methods attempt to go beyond counting words by parsing the sentence, which provides deeper insight into the meaning of words, but which relies on the entire sentence (or phrase) being uncorrupted. Information Extraction methods could be used, for example using regular expressions to extract the content of lists and tables, but again these methods would be more vulnerable to statistical noise than simpler statistical analysis. Both NLP and IE can be used on noisy data, but generally require the size of the corpus to be very large (much larger than in the present corpus) and redundant, so that redundancy can be used to compensate for a lower yield-rate of the methods. In the case of the present study, we opted for a simpler statistical study, with the belief that future work on more advanced algorithms will be able to extract higher quality information from this corpus. The statistical information that we extract in this study includes relative frequencies of words and phrases and their associated time series.

2.2 - Slovenian Newspapers

We additionally added three Slovenian-language newspapers from Gorizia to perform comparisons. *Soča*, *Gorica*, and *Primorski List* were digitised by the National and University Library of Slovenia, and are available online. They correspond respectively to 3342, 1472 and 892 issues or 38,153,317, 16,189,845, and 9,662,362 words.

The Slovenian-language newspapers from Gorizia were part of the documents digitised by the EU research project IMPACT ("IMProving ACcess to Text") (Impact 2008). Text was extracted from these newspapers using Abbyy FineReader 10, reporting CER of between 15-30% depending on the orthography used in the image (Jerele et al. 2011), and was made available by the Digital Library of Slovenia under a public domain license[8]. We only made use of the digital text already extracted and made available on the library website.

2.3 - Statistical Analysis of Textual Time Series

Trends, changes and continuities in the salience of various topics in historical corpora can be estimated based on changes in the frequency with which

carefully chosen words are used (Allen, Waldstein and Zhu, 2008; Lansdall-Welfare et al. 2017; Michel et al. 2011; Nicholson 2012). Our analysis of word frequencies and time series relies on the estimation of relative frequencies of words and phrases (their frequency in a time interval relative to the total number of words published in the same interval). To do so, we divided the period under investigation into three-month long time intervals (trimesters), and then computed the raw frequencies of each word and phrase, as well as the total amount of words used in each given time interval (volume).

As both Italian and Slovenian are highly inflected languages, the same concept can be represented by many different words. The standard solution to this problem is that of 'stemming' each word, removing the inflection and keeping a stable word stem, by means of a standardised procedure. We used the Snowball stemmer for both languages[9], which, for example, transforms the Italian words (*parola*, *parole*) into (*parol*) and the corresponding Slovenian words (*beseda*, *besede*) into (*besed*)[10].

After stemming, the text was split into n-grams up to a length of three, discarding all n-grams with a frequency less than 10. This was performed because another source of ambiguity, besides inflection, is OCR noise: errors made by the OCR software (particularly when operating on historical newspaper pages). This was dealt with by removing all n-grams that have been seen only a handful of times, since most errors will result in non-existent words, and are therefore either the result of corruption in the data acquisition pipeline, or are just so rare that they would not be of much interest in a study of large statistical trends. The resulting lexica sizes for each corpus can be seen in Table 3.

Table 3 - Lexicon size (number of unique n-grams) for each of the five corpora.

Corpus	**1-grams**	**2-grams**	**3-grams**	**Total**
Gorica	53,683	166,412	93,273	313,368
Primorski list	25,891	111,519	66,422	203,832
Soča	90,425	363,337	240,266	694,028
EDL	90,786	395,821	234,248	720,855
CFG	116,793	492,933	316,155	925,881

For each of these n-grams, we generated a time series of relative frequencies as follows. An important statistical consideration is that we do not have sufficient data to estimate the relative frequency of each word every day, in fact not even every month. This is due to the well-known property of natural languages (Powers 1998), that most words have a low probability of occurrence, and therefore to reliably estimate these rare words requires a large sample of text, which is larger than the daily (or even monthly) volume generated by any of the local newspapers considered in this study. As such, we resorted to estimating the relative frequency time series in time intervals that are three months long, thereby increasing the volume of data available to estimate the word frequencies at each time interval. This has the extra advantage that it removes the need to correctly segment the text into articles (a difficult and error-prone step) and to be accurate in applying dates to each of the digitised pages (a step that we did by hand). As we merge three-months content into a single time point, for which word frequencies are computed, these various sources of error are minimised.

Our processing pipeline can be thought of as estimating the probability of a word being used in each three-month period. The noise level of the OCR step has been estimated, as described in the previous section. Even though a word cannot necessarily be correctly detected every time it occurs, the probability of occurrence can still be estimated (much like the bias of a coin can be estimated from a finite sample of coin tosses).

Therefore, we represent the salience of a word (or phrase) over time by a time series formed by estimates of the relative frequency of each of the n-grams[11] during a trimester. This phase left us with 2,857,964 time series representing an estimate of the relative frequency of each n-gram in the newspapers within each three-month interval. In the rest of the paper, we will use these time series as a way to understand trends, changes and continuities in the aggregated newspaper content of Gorizia over 41 years.

In conclusion, the collection of 42 microfilms of the two Italian newspapers yielded 47,466 pages and 207,579 time series of individual words, representing their relative frequency in 168 (EDL) and 122 (CFG) time intervals of a trimester each**.** Also including the three Slovenian newspapers that we added to the analysis, the dataset that we built for this analysis is formed by 181,503,095 words and covers 164 trimesters from 1873 to 1914, as indicated in Table 4.

Table 4 - The number of time points available for each time series, where each time point represents a quarter (three months) in a given year.

Corpus	Total time points	First time point	Last time point
Gorica	59	1899 Q3	1914 Q1
Primorski list	84	1893 Q1	1913 Q4
Soča	171	1871 Q1	1915 Q1
EDL	168	1873 Q1	1914 Q4
CFG	122	1883 Q1	1914 Q3

2.4 - Statistical Overview by Word Frequency

As a first overview of the vast corpus, and as a sanity check of the processing pipeline, we looked at the most frequent words for each of the five newspapers. This is intended to help the Reader visualize what kind of tokens are produced by the data pipeline, what kind of errors can be found, and what kind of information one could extract from that. After removing stop words[12], we obtain the relative word frequencies shown in Table 5 (the most frequent words in each of the 5 newspapers, during the entire period under investigation). We note that the name of Gorizia / Gorica features very prominently, as expected, except for in the Catholic EDL, where the word *Chiesa* (Church) is one of the most common words, and that Trieste is present (in the two Italian outlets). The level of OCR error which is present in this type of statistical analysis can also be observed in the word frequencies.

3 - Textual Time Series Analysis

We are interested in seeing how a statistical analysis of this corpus can help us understand this crucial period for the city of Gorizia, the Habsburg Empire and Europe in general. Changes in the relative frequency of a word can reveal periods in which it was used more or less frequently, and hence reveal valuable information about what people were

Table 5 – Most frequent words in each of the five corpora, along with their relative frequency during the period of investigation. We note that the name of Gorizia / Gorica features very prominently, as expected, except for the Catholic EDL, where the word *Chiesa* (church) is one of the most frequent words. Notice also that Trieste is present (in the two Italian outlets).

Gorica	**Primorski list**	**Soča**	**EDL**	**CFG**
Gorici (1.0e-03)	Gorici (1.7e-03)	Gorici (1.6e-03)	tatto (4.4e-04)	Trieste (6.5e-04)
bode (8.4e-04)	katoli (8.1e-04)	kateri (7.6e-04)	Chiesa (4.2e-04)	Gorizia (6.2e-04)
poro (7.9e-04)	naro (7.7e-04)	bode (7.0e-04)	Trieste (3.8e-04)	signor (5.7e-04)
tern (7.1e-04)	bode (7.6e-04)	vseh (6.3e-04)	presso (3.7e-04)	presso (4.4e-04)
zbor (6.0e-04)	poro (7.4e-04)	zbor (6.1e-04)	tatti (3.6e-04)	tatto (3.9e-04)
gori (5.8e-04)	doma (7.1e-04)	torej (5.9e-04)	Gorizia (3.5e-04)	tatti (3.6e-04)
nega (5.6e-04)	gori (7.0e-04)	leta (5.6e-04)	Vienna (3.4e-04)	Vienna (3.6e-04)
niso (5.6e-04)	priporo (6.8e-04)	katero (5.5e-04)	venne (3.3e-04)	Giuseppe (3.4e-04)
leta (5.4e-04)	zbor (6.5e-04)	zopet (5.5e-04)	legge (3.3e-04)	Società (3.2e-04)
elni (5.2e-04)	niso (6.3e-04)	radi (5.3e-04)	cattolici (3.3e-04)	venne (3.2e-04)
torej (5.2e-04)	kega (6.2e-04)	treba (5.2e-04)	popolo (3.1e-04)	altre (2.7e-04)
zopet (5.1e-04)	leta (6.2e-04)	poro (5.1e-04)	gran (2.7e-04)	viene (2.6e-04)
kega (4.9e-04)	nedeljo (6.1e-04)	niso (5.1e-04)	altre (2.7e-04)	italiani (2.5e-04)
kateri (4.9e-04)	kateri (6.1e-04)	katere (5.0e-04)	ente (2.6e-04)	sera (2.5e-04)
ulica (4.8e-04)	ulica (5.7e-04)	toliko (4.9e-04)	partito (2.6e-04)	altra (2.5e-04)
vseh (4.8e-04)	posebno (5.7e-04)	naro (4.8e-04)	quei (2.5e-04)	quei (2.4e-04)
doma (4.7e-04)	zopet (5.5e-04)	gori (4.7e-04)	nome (2.4e-04)	Camera (2.4e-04)
vlada (4.6e-04)	ljudstvo (5.3e-04)	doma (4.7e-04)	numero (2.4e-04)	nome (2.4e-04)
Anton (4.6e-04)	torej (5.3e-04)	more (4.7e-04)	vero (2.4e-04)	corone (2.3e-04)
jako (4.5e-04)	vseh (5.2e-04)	vpra (4.4e-04)	Roma (2.3e-04)	pubblico (2.3e-04)
dela (4.5e-04)	elni (5.1e-04)	ulici (4.3e-04)	liberali (2.3e-04)	Antonio (2.3e-04)
vpra (4.5e-04)	nega (5.1e-04)	kega (4.2e-04)	Giuseppe (2.3e-04)	numero (2.3e-04)
krat (4.5e-04)	katero (5.0e-04)	imajo (4.2e-04)	altra (2.3e-04)	Consiglio (2.2e-04)
priporo (4.2e-04)	Anton (4.8e-04)	posebno (4.1e-04)	Camera (2.3e-04)	pare (2.2e-04)
katere (4.2e-04)	Gorica (4.8e-04)	strani (4.1e-04)	delta (2.2e-04)	fior (2.2e-04)
radi (4.2e-04)	katere (4.8e-04)	Gorica (4.1e-04)	signor (2.2e-04)	legge (2.2e-04)
tadi (4.2e-04)	vsem (4.7e-04)	Gori (4.1e-04)	parole (2.2e-04)	delta (2.1e-04)
katero (4.2e-04)	dela (4.7e-04)	svoj (4.0e-04)	viene (2.2e-04)	signori (2.1e-04)
posebno (4.0e-04)	radi (4.7e-04)	jako (4.0e-04)	pare (2.2e-04)	italiana (2.1e-04)
glede (4.0e-04)	treba (4.6e-04)	vsem (4.0e-04)	essa (2.2e-04)	Giovanni (2.1e-04)
more (4.0e-04)	Gori (4.6e-04)	nego (3.9e-04)	guerra (2.1e-04)	Francesco (2.0e-04)
pride (3.9e-04)	namre (4.6e-04)	Trstu (3.9e-04)	società (2.1e-04)	opera (2.0e-04)
niti (3.9e-04)	vpra (4.5e-04)	niti (3.9e-04)	cose (2.1e-04)	signora (2.0e-04)
treba (3.9e-04)	ulici (4.5e-04)	dela (3.9e-04)	mano (2.1e-04)	provinciale (1.9e-04)
toliko (3.9e-04)	nost (4.4e-04)	zato (3.9e-04)	buon (2.1e-04)	danni (1.9e-04)
svoj (3.8e-04)	mestu (4.4e-04)	nega (3.8e-04)	questione (2.1e-04)	parola (1.9e-04)
namre (3.8e-04)	zato (4.4e-04)	vlada (3.8e-04)	Italia (2.0e-04)	scuola (1.8e-04)
naro (3.7e-04)	imajo (4.3e-04)	pride (3.7e-04)	pubblica (2.0e-04)	lavori (1.8e-04)
izvr (3.6e-04)	toliko (4.3e-04)	krat (3.7e-04)	dato (1.9e-04)	Podestà (1.8e-04)
strani (3.6e-04)	duhov (4.3e-04)	dobi (3.7e-04)	fede (1.9e-04)	gran (1.8e-04)
zato (3.6e-04)	Tako (4.3e-04)	gotovo (3.7e-04)	parola (1.9e-04)	disse (1.8e-04)
Ivan (3.5e-04)	ampak (4.2e-04)	imel (3.7e-04)	diritto (1.9e-04)	Italia (1.8e-04)
nego (3.5e-04)	kron (4.0e-04)	Tako (3.7e-04)	cattolica (1.9e-04)	vero (1.8e-04)
dobi (3.5e-04)	delo (4.0e-04)	odbor (3.6e-04)	opera (1.9e-04)	Governo (1.8e-04)
vsled (3.4e-04)	Ivan (4.0e-04)	elni (3.6e-04)	delia (1.9e-04)	conto (1.8e-04)
Tako (3.4e-04)	izvr (4.0e-04)	svojega (3.5e-04)	giornali (1.9e-04)	parole (1.7e-04)
kron (3.4e-04)	jako (4.0e-04)	Slovenci (3.5e-04)	deputati (1.9e-04)	pubblica (1.7e-04)
imajo (3.4e-04)	imel (4.0e-04)	slovenski (3.4e-04)	cattolico (1.8e-04)	Comitato (1.7e-04)
nost (3.3e-04)	anje (3.9e-04)	ampak (3.4e-04)	Stato (1.8e-04)	italiano (1.7e-04)
vrste (3.3e-04)	more (3.9e-04)	nedeljo (3.3e-04)	Papa (1.8e-04)	festa (1.7e-04)

paying attention to, how they were responding to external events, what they were doing every day or even what their priorities were.

As this was a time of major cultural, political and technological change, and this land was a crossroads of different worlds, we have focussed on searching for the traces of this change, as reflected in this small border city. We have explored when technological innovation was adopted, when new political ideas were introduced, and how major external events were covered. We also look for more local information, about specific individuals and events, that are known to have been significant in the history or the economy of the city; technologies, ideologies and generally any words that can shed a light on the many changes that were taking place in those days.

It is important to note, at this point, that even 180 million words are considered small data for certain types of analysis, and so we cannot always expect a reliable signal from the estimation of words' relative frequencies. This is due to the fact that - in every language - most words have a low frequency, and therefore this frequency can only be accurately estimated on very large samples. The practical conclusion is that we should concentrate statistical analysis on the high frequency words and / or large changes, whereas for lower-frequency words we will use the statistical signals from the data as an inspiration for deeper explorations conducted with the classical historical method of close reading.

Ultimately, the best way to explore digital corpora is still an object of academic discussion, and the approach we follow is that of distant reading (Moretti 2013), or culturomics (Michel et al. 2011), that is extracting information about the general public discourse from statistical properties of the corpus, combined with a close-reading step, aimed at refining the keywords used, and providing historical context to the statistical information found in the corpus (Lansdall-Welfare et al. 2017; Nicholson 2012).

One possible first step of this approach to exploring digital corpora, however, is to seek the reflection of well-known events, trends or phenomena in the corpus, to understand to what extent the content of historical media reflects what is already known. This process acts as a sanity check, verifying that the data confirms existing widely-accepted 'ground truth' accounts obtained via independent means, such as the dates of specific events. We will then broaden our investigation, to see if the news content also supports what is believed or expected, or if it can shed new light on known events.

3.1 - Statistical Study 1: Four Known Events

We start by searching the data for the statistical footprint of four major events that affected the region, and the entire empire, during the time under investigation. Specifically: the Ljubljana earthquake of 1895, the Halley comet of 1910, the change of currency from Florin to Crown in the years preceding 1900 and the visit of the Emperor in 1900. These are events which are well-known and documented, with a clear date, thereby allowing us to test the hypothesis that a statistical analysis of this type can detect events covered in the news, as well as to calibrate our tools.

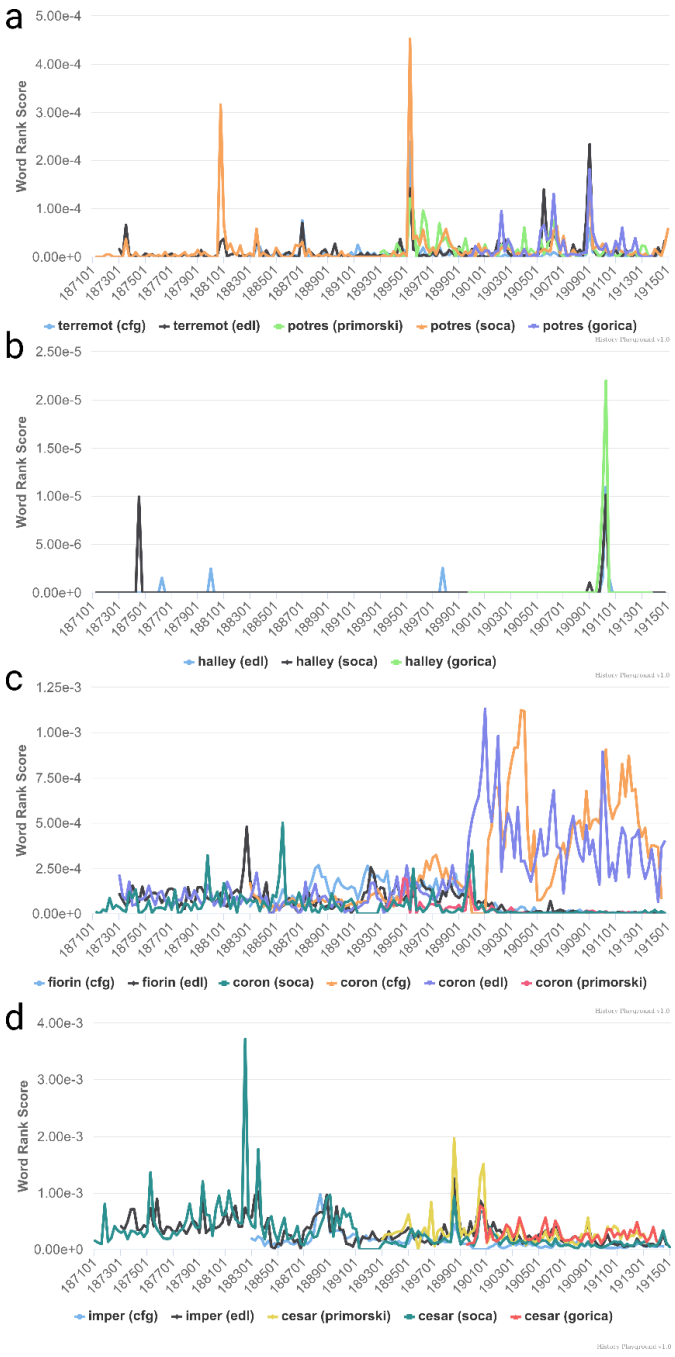

Figure 4 – The relative frequencies for known events were checked in the corpus. (**a**) Relative frequency of the words for 'earthquake' in Italian and Slovenian newspapers from Gorizia, between 1871 and 1914. (**b**) Relative frequency of the word 'Halley' between 1899 and 1913. (**c**) Relative frequency of the words corresponding to 'florin' and 'crown', the old and the new currency of the Austro-Hungarian Empire. The crown replaced the florin in 1900. (d) Relative frequency of the words for 'emperor' (It: imperatore, Sl: cesar).

Earthquakes. On the night of April 14th, 1895, a major earthquake struck the city of Ljubljana, today capital of Slovenia and then capital of a region called *Kranjska* or *Carniola*. The Slovenian centre lies 80 km east of Gorizia, with the earthquake being felt very clearly in Gorizia, and as far away as Vienna and Venice. Damage was reported in the territory of Gorizia, where the aftershocks continued for weeks. After five days *Corriere di Gorizia* was reporting that "among 900 houses in Ljubljana, nearly 700 were severely damaged and 300 are going to be demolished"[13].

By analysing the relative frequency of the word "earthquake" (It: *terremoto*; Sl: *potres*), shown in Figure 4(a), we can study the statistical footprints of that event in our collection, as well as the trace of some other major international earthquakes of that period: those of Belluno (1873), Zagreb (1880), Ischia (1883), Sanremo (1887), Calabria (1905), San Francisco (1906) and Messina (1908). We can also see that the Slovenian newspaper *Soča* responded more to the Slovenian quake, and the Italian *L'Eco del Litorale* responded more to the Italian earthquakes, suggesting perhaps a keener interest in Italian affairs for the Italian-language newspapers, and in south Slavic events for the Slovene newspapers. While this may seem obvious, it acts as an important gauge for assessing the reliability of the methodology, and of the corpora. This result provides evidence that the corpora are correctly fixed in time (with peaks occurring in the same place), and that the relative frequency comparison across corpora makes sense.

Halley's Comet. In 1910, two comets were visible in the sky, first the unexpected Great January

Comet, and then the much-anticipated May passage of Halley's comet. All newspapers in the world reported these sightings, and so these events provide a perfect signal for benchmarking and calibrating news mining tools.

In Gorizia, the passage of the comet was considered with suspicion by some people. Many feared it might be a harbinger of bad news, but the newspapers in our corpus mostly wrote about it in scientific terms.

The dismay and anxiety engendered by the passing of the comet epitomized a general sense of unease that characterized the decades before WW1 in Mitteleuropa. It pervaded art and culture, stimulating Freud in his discoveries, Kafka and Joyce (who spent many years near Gorizia, in Trieste) to develop their narratives. It also evoked the gloomy feeling of being on the brink of a great catastrophe: the social and ethnic equilibrium of the Austro-Hungarian empire was cracking, and new technologies were disrupting ancient habits. The continuous victories of man versus Nature seemed to be pointing to a revenge of the latter. A well-known poet from Gorizia, Michelstaedter, was one of the best representatives of these self-destructive tensions pervading the Habsburg society. He took his life in the October of that same 1910, following a dark culture of suicide which even involved the only son of the emperor almost twenty years earlier.

L'Eco del Litorale, in a small article entitled "The Fear for the Comet - The End of the World!" wrote about the terror of many Slovenian and Croatian farmers in the wider Littoral and, mentioning a Governmental report, wrote that:

"many peasants are thinking to sell all their properties and party until the end, taking it as it comes. It is a replay of the terror of the year 1000, but in a happier fashion".

The article continues that the authorities, concerned for these anxieties, "were asking school teachers and parish priests to inform people about the theory of comets", while a government agency was about to distribute a booklet on the topic in schools and churches[14].

In Figure 4(b), we can see that our corpus shows a clear sign of the passage of the comet, indicated by the relative frequency of the word "Halley" in the local press of Gorizia, the same word being used in both languages, providing us with a second calibration of the time series extracted from both Italian and Slovenian newspapers.

Crown, the new currency. The changing of a currency in a country strongly impacts all aspects of life, and so it should be easily seen in the content of newspapers. The Austrian-Hungarian empire shifted to the gold standard in 1892 and smoothly introduced the Crown in place of the Florin. The old coin ceased to have legal tender in 1900 (Cvrček 2013).

This smooth transition can be read in the data, with a decline of the word "*fiorino*" in the 1890s accompanied by a growth of the word "*corona*". However, we encountered some difficulties in finding sufficiently reliable data in the Slovenian newspapers, since they mostly used the contracted versions *gl.* and *k.* for the different currencies, making them difficult to distinguish from the rest of the data (these short strings, combined with OCR noise, can generate very ambiguous signals). Despite these impediments the turning point of 1900 is quite clear in Figure 4(c).

The Emperor. It is interesting to measure the extent of the coverage of emperor Franz Josef (It: *Francesco Giuseppe*; Sl: *Franc Josip*), who reigned over the Habsburg Empire from 1848 to 1916, and was an important figure both for those who appreciated him and those who were less supportive of the Empire. A lack of support could not be expressed in the media anyway, as nobody was willing to risk fines and jail in Gorizia, with the exception of a few young Italian liberals and their irregular press (De Grassi 1982, 57-8; Horel 2015, 89). Most of the common citizens in the Littoral considered him the highest institution in their country, having been on the throne for almost seventy years, probably engendering a certain sense of trust and affection towards him. Media coverage of Franz Josef included also the tragic death of his only son, that of his brother Maximilian (who was executed in Mexico in 1867) and of his wife, the empress Sissi, who was very popular, but was murdered in Switzerland in 1898. The assassination of his nephew Franz Ferdinand began WW1.

Franz Josef's visits to the region were always a major affair. He visited Gorizia five times: in 1850, 1857, 1875, 1882, and 1900. The last visit on September 29th, 1900 marked 400 years of Gorizia as part of the Empire and was prompted by the Italian liberal city council who formally invited him (Agostinetti 1981, 42). Those events, and visits to Gorizia covered by the time frame of this research left a peak of the word "emperor" (It: *imperatore*; Sl: *cesar*) in our dataset. Note that Franz Josef was coronated on December 2nd, 1848, and we can see peaks in Figure 4(d) in most of the following 10-year anniversaries.

3.2 - Statistical Study 2: The Second Industrial Revolution and the Belle Époque

The period between 1873 and 1914 was a time of great social change and saw the introduction of an extraordinary number of innovative technologies and ideas. While telecommunications and trains were already in use, they started to have a deeper impact on the region and its economy in this period. The telegraph was outpaced by the telephone for the first time, and the rail network greatly improved: in Gorizia, the first train station was opened in 1860, with the second station brought into operational use in 1906.

These were just two of the ways in which new people and new ideas poured into Gorizia: in the period under study we also see the introduction of cinema, airplanes, cars, gramophones, electricity, bicycles, and a plethora of new chemical products. Local newspapers, of course, became widespread. It was a time of great economic development, especially the last two decades, those of the so-called Belle Époque.

Besides technology, from the 1870s there was also a growth in social unrest, the emergence of socialism, the Catholic church adopting the 'social teaching' with the Rerum Novarum in the 1890s, and the general rise of nationalism. In the Habsburg Empire, there were political reforms, increased ethnic tensions, as well as ideological tensions. Industrialization of the land surrounding Gorizia set in motion an increase in immigration from the countryside to the urban centre, bringing an ethnic shift along with it. These were all signs of the second industrial revolution and of a society at large moving from a feudal-agriculture to a capital-

industrial order, where the masses acquired new roles in economics and politics.

Old and new technologies and habits coexisted in the same environment; public opinion gave voice to the fears and worries of the time, mixing them with the political confrontation of the day. The following lines were written in EDL about the opening of the second station in 1906, where we hear about Italian Catholics' complaints about the Italian national-liberals, who ruled the municipality of Gorizia, encompassing a request for modernization of local transportation, criticism of city planning, and the ubiquitous obsession for national rights:

"Have you ever seen those moving shacks [...] that are the city trams? [...] "Din, din, din": so the carriages go with their wobbly roof. They are pulled by nags that seem to have been cursed. They advance head down like a beaten dog, skinny, stiff, hardly moving their legs. [...] at the same time we were thinking about how the things at the municipality are in a bad shape, so badly managed. The movement of travellers is going to increase. From the Transalpina station and the Southern station. [...] where are those travellers going to be placed? Will the city managers have the courage to pack them into those tram carriages [...]? What about the project for the new tram? [...] but unfortunately at the city council they scream just for rights, either presumed or founded, without recognising that there are some duties to accomplish too."[15]

Transalpina Station. The train was one of the main drivers of the modern age and of industrialization, connecting distant places and bringing the age of

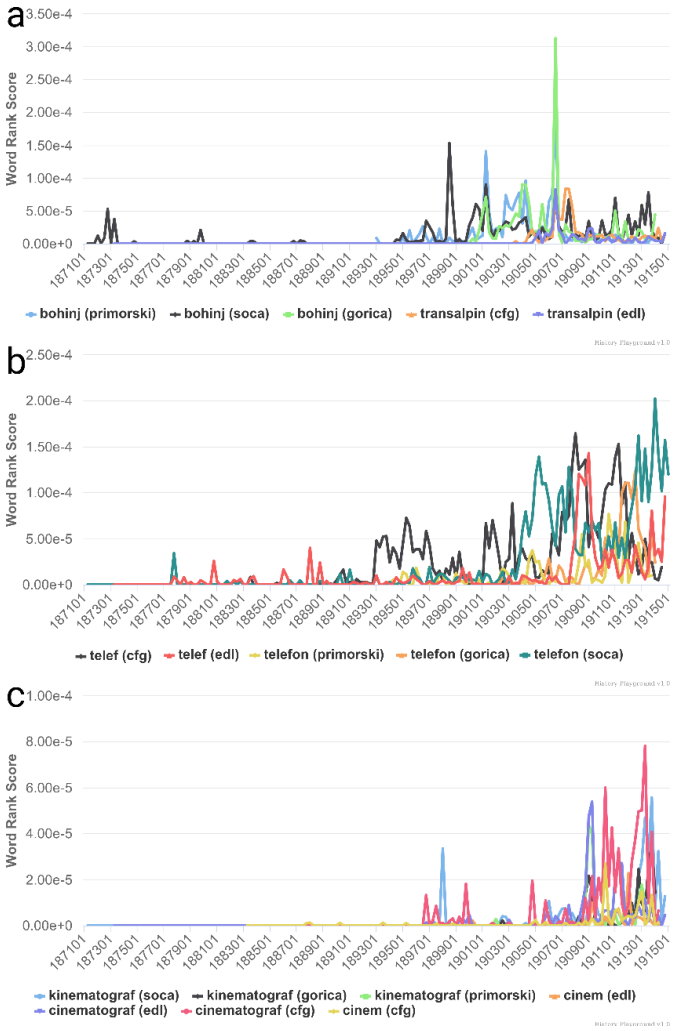

Figure 5 – **(a)** The relative frequency of *Transalpina* and *Bohinjska*, names that are used in reference to the new train line (and station) of Gorizia. *Bohinjska proga* was the Slovenian name of the railway line that the Italians called *Linea Transalpina*. **(b)** Relative frequency of the words *telefono* and *telefon* in Italian and Slovenian newspapers, from 1871 to 1915. **(c)** Relative frequency of the stems *kinematograf-*, *cinematograf-* and *cinem-* from 1871 and 1915.

machines into everyday life. In Gorizia, the first line, the "Southern", arrived in 1860 (Inoue 2008, 8). It linked the city with the capital Vienna and the main port of the empire, Trieste. It brought imported products from all over the world and tourists from

the colder inner regions of the monarchy. Together with a new station, the new line, the "*Transalpina*" (known as *Bohinjska proga* by the Slovenes), opened in 1906 (Inoue 2008, 64); it again ran North/South but on a more direct path, while offering cheaper tickets for goods and passengers. The local press was present at its opening, after witnessing its last stages of construction, and again later, after the line was fully operational. Figure 5(a) shows the timeline of the relative words, peaking exactly in 1906, and remaining in constant use afterwards, showing how it became a part of daily life and the economy from its very inauguration.

Telephone. Local public opinion was fascinated by the invention of the telephone since the first experiments in the 1870s. *L'Eco del Litorale* poetically refers to it as the "speaking light" in 1880[16]. However, when reporting the achievements in transatlantic communication, *Corriere di Gorizia*, by using humour of its epoch, would sound quite sexist in our age:

"From now on it is going to be possible to have a conversation at a 2600 miles distance, speaking 120 words per minute, quite a speed for female tongues."[17]

Despite all the enthusiasm for this new invention, the telephone was introduced relatively late to Gorizia, in 1894[18]. It was installed in two public places at the beginning: at the central post office, and at the post office of the train station, while twenty subscription requests were placed before the service was even finalized by authorities[19]. However, in our corpus, we find that telephone was embraced much later by the general public: mentions of the telephone only became widespread in the newspapers in the following decade, related to the increased publication of advertisements in the newspapers, another trademark of the Belle Époque. Figure 5(b) shows the 1894 start of *telefono* in the Italian newspapers and *telefon* in the Slovenian ones, and the following steady growth in frequency of both words.

Cinema. The first cinema projection in Gorizia took place on December 8th, 1896, at the Hotel Central, with the first purpose-built cinema, Cinema Edison, being opened by Andrea Kumar 12 years later in 1908. On the same site as that first cinema projection, a permanent movie-theatre was later opened in 1909, by Josip Medved, and named Central Bio, derived from "bioscope", an early word for cinema. Trilingual programs were printed for all its screenings, showing how all ethnic components and social groups were included by this new medium. In those years between the first projection and the first permanent cinemas, travelling shows regularly visited the town for the seasonal fairs, bringing the latest films with them. Spectacles were held many times per day, and in the night, there would even be "gentlemen only" shows[20]. Silent cinema was the popular, modern and cheap attraction that signalled the rise of mass society. It was a show that everybody could understand, quite removed from the opera which was a pastime more suitable for the educated middle and upper classes.

On November 24th, 1896 *Corriere di Gorizia* writes on its front page an extensive technical description of a cinematograph, entitled "Animated photography*":*

"It is an absolute novelty that - for those who do not know its mechanism and physical laws - appears to be a marvel."[21]

Then on December 10th, 1896 the same journal reports on the first projection in Gorizia, describing the famous scene of the arriving train that quickly became part of the world's collective memory:

"The views of this cinematography are various: we see […], a woman bathing in the sea, the arrival of a train, the movement of the passengers. This railroad scene was the most interesting: one sees the arriving convoy, the conductors opening the doors, the passengers descending, some picking up luggage, some their dog, everything very clear and very good, so much so that this section was especially applauded".[22]

Figure 5(c) shows clearly the time of the first show in 1896, the various visits of traveling cinemas, and the early ill-fated attempts to establish a permanent cinema in town until in 1908, when several permanent cinemas become part of everyday life in Gorizia and in its newspapers. This shows the extent to which this mass medium became adopted as a part of the city life, in both ethnic groups.

Airplanes. Included among the many movies shown in local cinemas were the accounts of Scott's expedition to the South Pole, and of the tragedy of the Titanic. In the autumn of 1909, a film was aired in Gorizia about Louis Bleriot's crossing of the Channel by airplane. This cannot have been missed by two young brothers, Edvard and Josip Rusjan, bicycle mechanics who had been building gliders and trying to build a full airplane, just like the Wright brothers (Scandolara 2001, 11). Edvard would buy the same engine as Bleriot (an Anzani 25hp) and by November 25th, he had completed the first powered flight in the Habsburg Empire, flying first for 60 meters, then on November 29th flying for 600m, and finally flying over Gorizia. He eventually died flying over Belgrade in January 1911 during a flight demonstration. The story of these two brothers exemplifies the complexities of this border territory and this time of transition: their father was a Slovene from Trieste, their mother a Friulian from Medea, but they referred to the first glider (made from bamboo and paper) with a Venetian expression: "*La Trapola de Carta"*[23] (the paper contraption) (Mlakar and Turel 2010, 145-6).

The local media covered aviation in general, but with a special interest in the Rusjan brothers. We use the Italian word *aeroplano* and the Slovenian *legato* in Figure 6, noting that *Soča* appears to use the Italian *aeroplano* too. The Slovenian newspapers covered these inventors more than the Italian papers, even after we account for alternative possible spellings of their name (we tested "Rusjan", "Rusian" and "Russian", the last being a common name in the area). What is equally important is that the local media showed interest for this innovation from its start: just seven years after the first flight of the Wright brothers, the town of Gorizia had seen its first flight of a local-made aircraft.

Electric Lighting, Bicycles and Cars. The town where the Rusjan brothers built their airplanes was as modern as their bold flying exploits. The years leading up to that first moment of powered flight had seen a remarkable rate of social and technological change, which galvanized Gorizia as

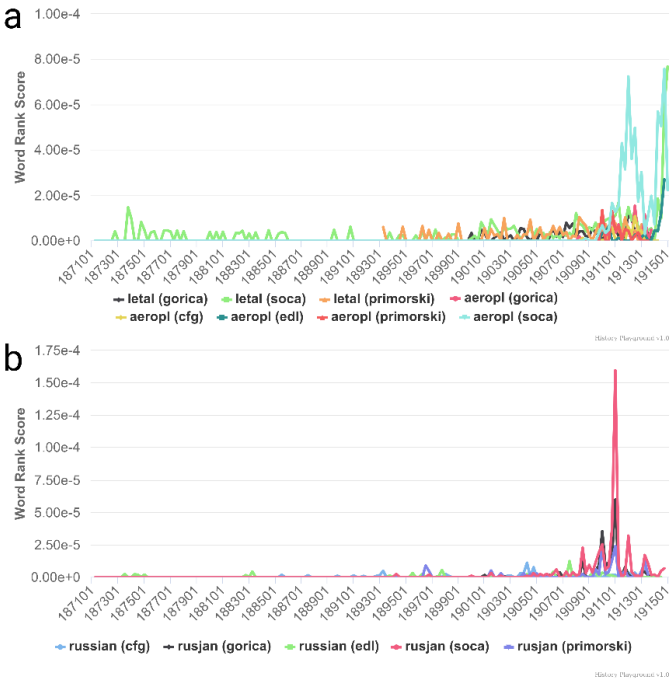

Figure 6 – **(a)** Relative frequency of the stems corresponding to *aeroplano* and *letalo* (airplane). **(b)** Relative frequency of the words *Rusjan* and *Russian* (alternative spellings of the surname of the flight pioneers from Gorizia).

it did the rest of Europe, but it also stirred up some anxieties. The turn of the century did not just see the introduction of the telephone and cinema, and a change of currency, but also the adoption of other key technologies, such as electric lighting, bicycles and automobiles.

While gas lighting in the streets had been introduced in 1871, from the late 1890s tests were being conducted to deploy electric lighting, which were eventually introduced in 1903. The newspapers reported some complaints about this innovation, but also printed admiring descriptions of the new electrical plant that was to serve the city: a 115 horsepower gas engine coupled with a dynamo to produce a current of 480 volts and 142 amperes was installed in early 1903[24] in preparation for the August 15th start of the new service[25].

The first automobile was seen in town in 1898 and promptly reported by the newspapers: "Yesterday an automobile was wandering around our town, attracting everybody's gaze and attention. It was ridden by two brothers, the counts Gyulay, travelling from Vienna [...] This automobile covers one kilometre in 2 minutes on flat terrain"[26].

Bicycles (or velocipedes) had become available from the 1860s and had encountered early success, with the 1890s seeing a real "bicycle boom". They were embraced by Gorizia during this boom, where bicycle races became commonplace. An interesting race between a car and a bicycle is reported in local newspapers in 1901, won by the bicycle with a margin of five meters[27]. An 1898 letter complaining about reckless cyclists in town said:

"Every day we deplore serious events, the other day in Piazzutta a girl was nearly the victim of the reckless speeding of a velocipedist, two days ago it was the turn of a lady in the Corso, and this morning a tragedy happened, which nearly led to irreparable consequences. [...] These 'brave' people take off rather than stopping and showing interest for their victims."[28]

The adoption of these three technologies from the second half of the 1890s can be seen in Figure 7, where we compare the words for "automobile", those for "electric" (not just lighting), and we only show the Italian word for "bicycle" because the Slovenian word *kolo* is very ambiguous (referring also to any wheel, and to a dance).

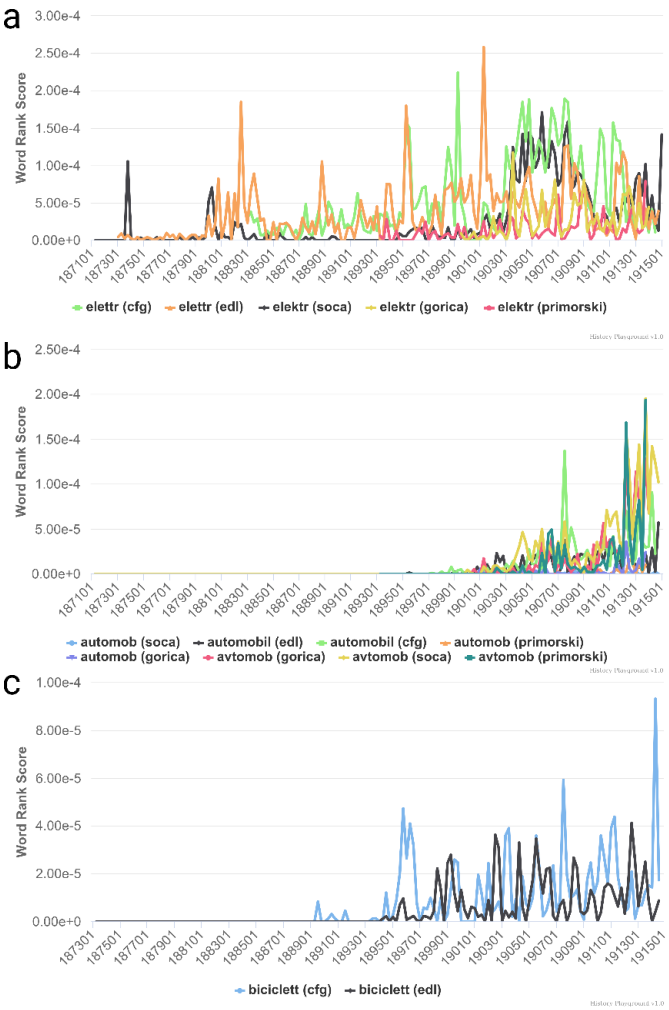

Figure 7 - The relative frequency of **(a)** 'automobile', **(b)** 'electricity' and **(c)** 'bicycle'. We see that the end of the 1890s is a period of fast adoption of new technologies.

The Lunatic Asylum - The 19th c. was an age of increasing rational management of most aspects of life, with health management being no exception. The modern hospital was a protagonist in the medicine of the late 19th c., while mental health, which was particularly developed in the Habsburg empire, also needed its own dedicated sites.

The debate around a provincial lunatic asylum enveloped the city from the end of 1880s. Its construction was officially announced by authorities in 1891, and in 1913 the *manicomio*, as the Italians called it, was opened, just to be

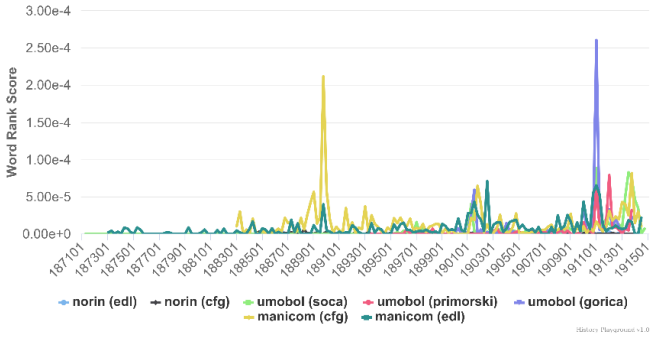

Figure 8 - Relative frequency of various words relative to mental asylum: *manicomio*, *umobolnica*, *norišnica*.

destroyed a couple of year later by the war (Fabi 1991, 77-8, 82-3). We can see that the months preceding these announcements saw an increase in debate and that there was a shift in the words that Slovenians used to call it. From using *norišnica* up to the beginning of the 1890s, they dramatically moved to using *umobolnica*.

Nonetheless, any issue debated in the public arena was framed within the national conflict between the Italians and Slovenians. These words from *Corriere di Gorizia* in 1889, a time when debate was peaking about the lunatic asylum, demonstrates this point clearly:

"[...] it is deplorable the incongruence of some councilmen from the Slavic side, who did their best to make the urgent establishment of the Lunatic Asylum for our province an illusory achievement. [...] in the issue of the Lunatic Asylum there was a furious quarrel between the councilmen of the two nationalities, like that in Babel, being convinced that for the time being any argumentation on this trite issue has become pointless."[29]

Figure 8 shows peaks preceding the completion of the hospital in 1913, and before the authorities officially resolved to build it in 1891,

showing the increased debate covered in the newspapers at the time.

3.3 - Statistical Study 3: Social, Political and Cultural History

A significant portion of the news in these newspapers was devoted to issues of politics and nationality. Even issues that would not naturally be framed in this way were sometimes reported within a discourse of national and political conflict.

Together with technological, social and economic shifts, the world described by our dataset seemed to be experiencing a change in the very texture of how its collective identities were represented, and how various social groups were perceived. The rise of socialism, and of the Catholic Social teaching, introduced an awareness of class divisions, and even of the possibility of a conflict among them. In a region where city and countryside, Italians and Slovenians, upper, middle and lower classes had previously occupied the same position for centuries, the time of change led by the second industrial revolution and the dawn of mass society proved disruptive and destabilizing. In that period, we can see the formation of the national-ideological nexus that deeply influenced the politics of the North Adriatic border region in the 20th c. and that we discussed earlier in the paper. We now look at some of these topics and, aware that they are among the hardest to be understood in quantitative terms, test our methods and report their outcomes.

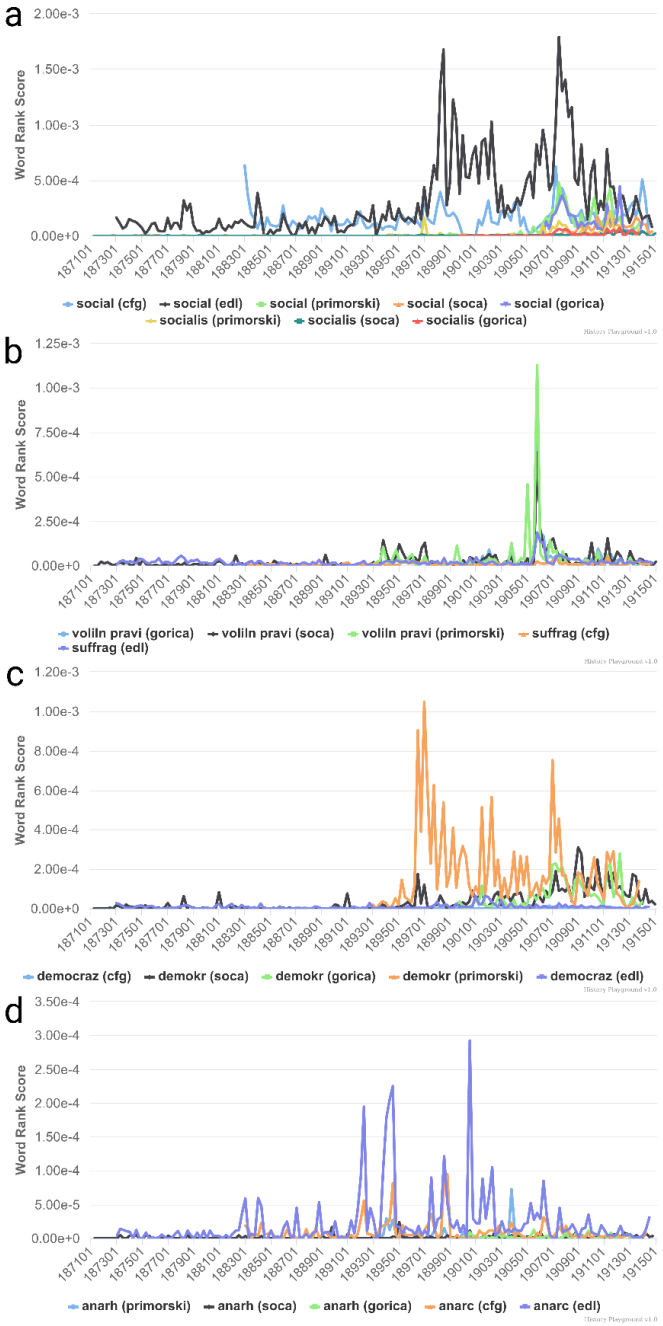

Figure 9 - The relative frequency of the word stems relative (**a)** to socialism (in Italian and in Slovenian) show a rapid increase at the end of the nineteenth century, **(b)** to suffrage (It: *suffragio;* Sl: *volilna pravica*), **(c)** to democracy change frequency after 1897 and **(d)** to anarchy showing peaks corresponding with high profile murders.

A social and ethnic road towards the new politics. The Socialist party of the County was born in 1902, after an alliance between the first scattered groups of factory workers in the area had emerged a few years earlier; nonetheless socialism never achieved particular prominence in the city of Gorizia (Patat 2003, 79). Social issues had risen to prominence, in the wake of the many

transformations that were under way, but they were promptly acknowledged by the Church, which developed a Christian Social doctrine after the publication of the Rerum Novarum encyclical in 1891. This meant that the clergy provided social services, while still attacking secular thought, a strategy that paid off - for example - at the first county elections with universal male suffrage of 1907, won by the Christian Socialists (Agostinetti 1981, 50-1). We can see the Catholic perspective on the social question, by close reading of EDL:

"Father Pavisich states that the social question exists because, whether we like it or not, there are infinite riches in the hands of the few, and misery, poverty, hunger for the many, today in the world. Using statistical data he demonstrates what really is the impoverishment of the people provoked by capitalists' selfishness. [...] Father Pavisich believes that the primary cause of this intolerable situation is the French Revolution. [...] But now the people are revolting against this situation, massively following the red flag of democracy, thus creating a harder conflict, that preannounces a world catastrophe."[30]

In Figure 9(a) we can see that discourse about social issues and socialism had been increasing from approximately 1897, peaking around the dates of the various local elections (which took place in 1897, 1902, 1907, 1911, 1913). We can also see that EDL shows a particular interest in social topics. A great expansion in voting rights occurred in 1897, with universal male suffrage starting in 1907, leaving a footprint of the important debate preceding the actual vote shown in Figure 9(b). While there was a restriction on the number of representatives for low-income voters, the new system did give a new voice to the lower classes, translating into more influence for the Slovenian groups. In the same period, we can also see that discussions of democracy increase in frequency, as shown by Figure 9(c).

Anarchism had even less popularity than socialism in the County of Gorizia and Gradisca, but its appearance in the international political scene concerned Gorizia's journals, especially when lone anarchist attackers killed prominent figures and the press was eager to cover the inevitable trials that ensued. The murder of the French president in 1894 resonated more than any other similar event in our press, but the homicide of the popular Franz Josef's wife Sissi, in 1898, and of the Italian King Umberto in 1900 can also be seen in the time series of Figure 9(d).

Political quarrels and ideological debate in the newspapers of our dataset, especially relating to election seasons, reflected (and possibly also shaped) reciprocal perceptions of the two ethnic groups. Historiography on Gorizia remarks that in the few decades before WW1 there was also an increase in national conflict, with one group often pointing at the other as a competitor, especially with the universal suffrage of 1907 boosting the ethnic conflict (Fabi 1991, 24-5; Ferrari 2002, 367; Kacin-Wohinz and Troha 2000, 69-79; Marušič 2005, 317-44).

From the data, we can confirm a growth in references to Slovenians in the Italian newspapers, however the Italian journalists often used to refer to Slovenians as Slavs, a more generic, possibly slightly disrespectful term. The Italian national-liberal *Corriere di Gorizia*, while increasingly attacking Slovenes in the 1890s, far preferred the term *sloveni* to *slavi*. After it was silenced by

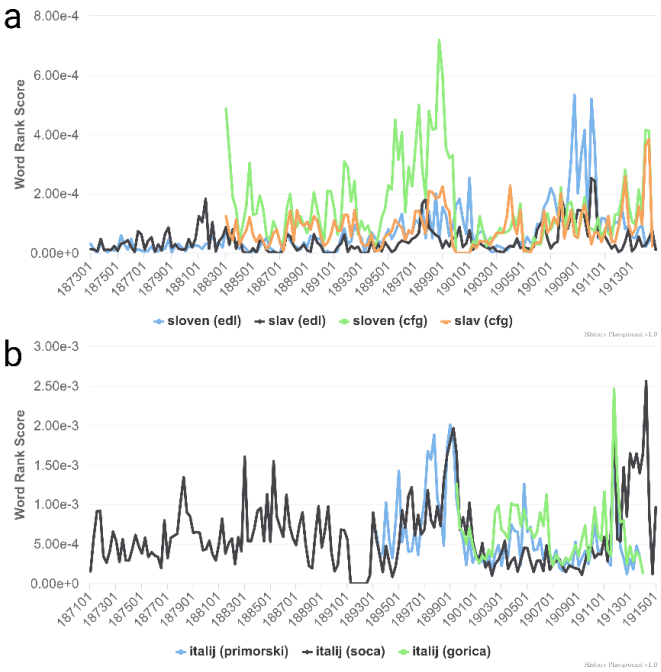

Figure 10 - The prevalence of ethnic discourse, represented by mentions of word stems for 'Italian' and 'Slovenian', seems to increase from the last decade of the nineteenth century.

authorities in 1899 and reopened in 1901 as *Il Corriere Friulano*, the term *slavi* received almost the same frequency as *sloveni*. Slovenian newspapers in turn incremented references to Italians in the years of this study. Altogether, Figure 10 shows an increase in political and ethnic awareness in the newspapers of Gorizia, mostly starting in the last few years of the nineteenth century. Another increase of tension seems to be visible in the years before the war.

1879-1890 Taaffe's Governments. Some of the clearest signals that we can find with our methodology are the statistical footprints of major political events at a national level. During Eduard Taaffe's government in Vienna, from 1879 to 1893, the central powers recognized the irreversible process of collective national identities building, and constantly looked to maintain an equilibrium among the various nationalities through concessions and specific laws. Taaffe tried to use the institution of the monarchy as a glue to bind together an imperial collective identity (Ferrari 2002, 340-1).

The Catholic newspaper EDL, soon after Taaffe came into power, recognised the prevalent liberal tendencies of his incoming government. However, it hoped that his politics and the wide coalition that supported him, officially beyond any nationality and ideology, "would deliver for the future a fair, impartial administration, not benefiting one ethnicity over the others"[31]. Taaffe affirmed being "above parties"[32] and, at the beginning of his mandate, clearly "set a target in the grandeur of the empire and the happiness of nationalities who are a part of it, willing to satisfy their legitimate aspirations"[33].

It was an important step for the central government to talk about the legitimate aspirations of the nationalities that are part of the empire. In the long run, Taaffe's politics proved ineffective, and national groups conflicted more and more, weakening the fabric of the Empire (Ferrari 2002, 340-1). In Figure 11 we can see Taaffe's footprint in Gorizia's newspapers clearly traces the fortune of his government.

Italian and Slovenian organisations. The two ethnic communities of Gorizia found different ways to associate and organise, initially around sport, and later also around cultural and political societies. The Italians created the *Unione Ginnastica*, and the Slovenians had *Sokol*, both devoted to sports and originating before the constitutional reforms of

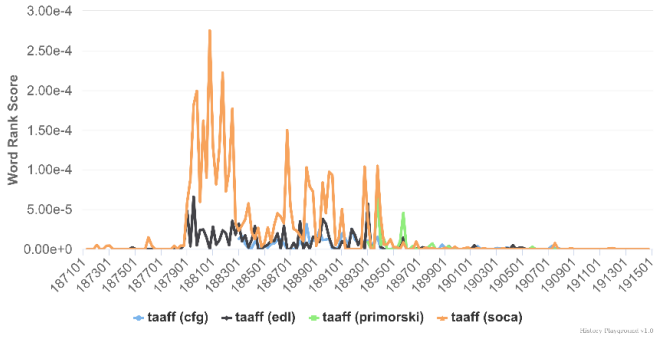

Figure 11 - This time series of the word 'Taaffe' shows the period in which he has been active as Prime Minister.

1867. The 1870s saw the creation of the *Sloga* society to coordinate the Catholic and liberal components of the Slovenian community (which soon made the newspaper *Soča* its mouthpiece), while the Italian component relied on more informal cultural circles. The 1880s saw the creation of cultural societies, including for the promotion of schooling in the national languages: *Pro Patria* for the Italians, which would later turn into *Lega Nazionale*, and *Ciril in Metod* for the Slovenes.

"If the Germans in Austria have Schulverein, and if the Slavs have their Cyril and Methodius society, both societies born not just to protect, but to propagate (I said propagate) their two languages, why on earth cannot we Italians have a Pro Patria society to protect and cultivate our sweet language, despite being subjects of the same State of theirs, and so having their same rights?"[34]

We found these words in an article appearing in 1887 in *Corriere di Gorizia*. They confirm established historical interpretations on the rise of national organizations in Europe. In his masterpiece, *The Nationalization of the Masses*, George Mosse clearly shows how national cultural associations begun spreading in Germanic countries in mid-19th c. The *Deutscher Schulverein* was founded in Austria in 1880 and provided basic schooling, framed in a national context, promoted German culture through public events, and printed journals and books. They were soon followed by similar organizations, first in Bohemia and then among other regions inhabited by Slavic populations, such as Slovenia or the Littoral, who opened new schools operating in their national languages (Mosse 1975).

The *Cyril and Methodius* society (*Družba svetega Cirila in svetega Metoda*, or simply, *Ciril in Metod*) was founded in Ljubljana in 1885 and soon spread to the surrounding areas. The popularity of Cyril and Methodius spread greatly in all Slavic countries from 1880, when Pope Leo XIII, in the encyclical *Grande Munus* (Great Duty), officially praised the two saints as the first Christianizers of Slavic populations in the high middle ages (Filipič 2010).

Italians in the Habsburg empire followed the same mitteleuropean template, taking a different path from those in the Kingdom of Italy, who did not rely on these types of societies. *Pro Patria* was founded almost simultaneously in Gorizia, Trieste, Trento and other major Italian Habsburg cities in 1885. This competition in the cultural field saw Italians and Slovenians matching each other's moves. Vienna accused *Pro Patria* of being a supporter of irredentism and outlawed the Italian organization in 1890; nonetheless it soon reopened as *Lega Nazionale* (National League) (Ferrari 2002, 356-7; Redivo 2005, 19-36). Even before the constitutional reforms of 1867 opened the door to the right to form real associations, sports societies were used as the *de facto* organisations through which national and cultural identity was propagated,

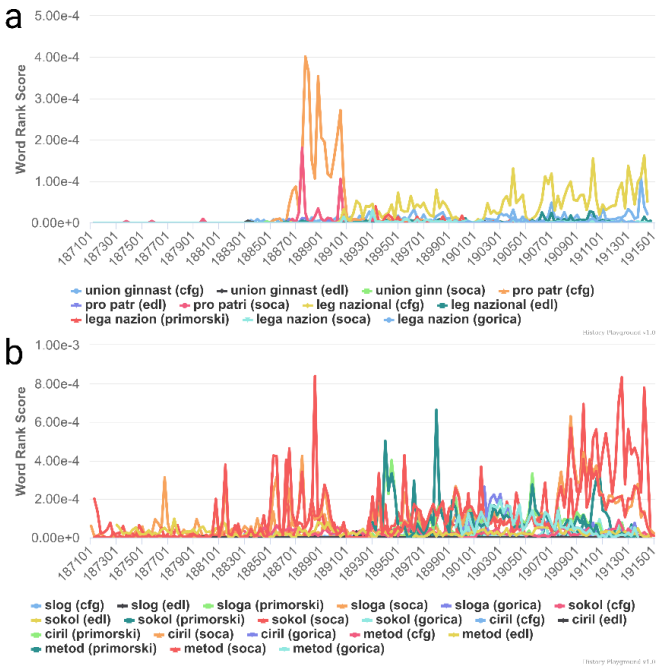

Figure 12 - The word stems relative to some of the key national associations in Gorizia: **(a)** *Lega nazionale*, *Unione Ginnastica*, *Pro Patria*; **(b)** *Ciril* in *Metoda*, *Sokol*, *Sloga*.

turning sporting activity into a way to express national cultures within the empire. Again, following the German national model, from the beginning of the 1860s Italian and Slovenian associations began to spread in the Habsburg empire. The first of them were the *Ginnastica Triestina*, in Trieste, and the *Gimnastično društvo Južni Sokol*, in Ljubljana, both founded in 1863, one year after the first *Sokol* began in Prague. In Gorizia, the *Unione Ginnastica Goriziana* was established in 1868, while the *Sokol* was later established in 1887.

The emergence of all these association in our corpus is visible if we look at the plots in Figure 12. The peak in 1880-1 for Cyril and Methodius confirm the consideration of the *Grande Munus* as well.

4 - Geographical Bias

As we have previously covered, the County of Gorizia and Gradisca was formed by various geographical regions, each with a slightly different history and ethnic make-up. For example, the area by the sea used to belong to the Republic of Venice until the Napoleonic wars, the area east and north of Gorizia spoke mostly Slovenian, and the area west of the Isonzo and Gorizia spoke mostly Friulian. In this section, we consider which geographical locations of the County were most mentioned in each of the newspapers.

In order to have an unbiased study, we started from a list of local communities, listed in the cadastre registers of the County at the time using, wherever possible, both the Italian and Slovenian names for each location. The list of names used comes from the *Franziszeischer Kataster* (It: "*Catasto Franceschino*", Sl: "*Franciscejski kataster*") at the *Archivio di Stato di Trieste*[35]. For each location, we endeavoured to find translations in each language that are appropriate to the period. However, we should note that in a few cases, some words might have their count inflated by synonymy (e.g. the word *Trenta*, which is a valley in Slovenia, is also the numeral thirty in Italian).

For each of these locations, we counted how often each was mentioned in either the Italian or Slovenian newspapers, normalised by the total number of locations mentioned in each corpus and visualised the results by placing markers on a map, sized according to how often it was mentioned in the Italian or Slovenian news.

The overall pattern that emerges is clear and shows how newspapers of each nationality have a distinct geographical focus, shown in Figure 13. We

see that Italian newspapers mention more frequently locations from the coastal south (Grado, Monfalcone, Aquileia, etc.) and Friulian east (Cormons, Caprica, Villesse, etc.); while the Slovenian newspapers mention more frequently locations from the Slovenian north (Tolmin, Kobarid, Bovec, Solkan) and Karstic east and south-east (Komen, Vilje, etc.). This geographical pattern in the choice of locations mentioned in the news of the time reflects both the older borders of the

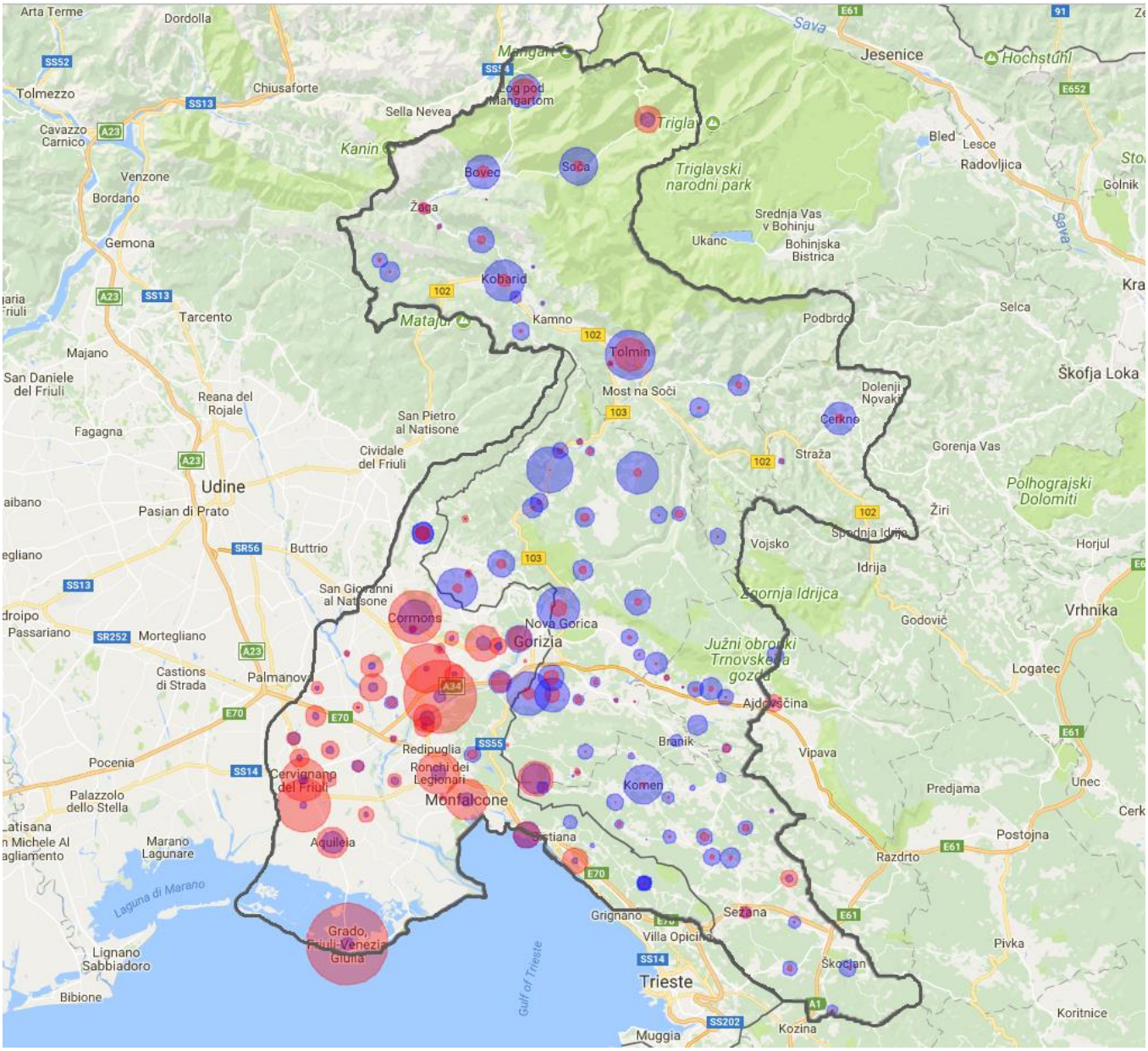

Figure 13 - Each circle on the map shows the proportion of times that a given geographical location of the County is mentioned, out of all mentions of County locations in that language (red for Italian, blue for Slovenian). We see that Slovenian newspapers tend to mention more locations to the north and the east of Gorizia, Italian newspapers pay more attention to locations to the south and the west. These areas reflect both the older boundaries of the Republic of Venice (in the south), and generally the traditional regions of Friulian speakers (plains east of the river), Venetian speakers (coastal plains and lagoon) and Slovenian speakers (Karst, Soca valley and Vipava Valley).

Republic of Venice and the traditional geographical regions inhibited by Friulian, Venetian and Slovenian speakers. In this case, the placement of each newspaper's core readerships is a likely explanation for the differences in local geographical news coverage.

5 – Conclusions

In recent years, many studies have focused on the analysis of historical newspapers. A recent article by Franzosi entered into a debate about historical methodology, directly addressing the 1983 debate between Fogel and Elton (Fogel and Elton 1983) - which centred on two different ways to understand the past – by proposing a third road to the past, one that draws heavily on the computer-assisted analysis of historical sources like newspapers (Franzosi 2017). Previous work by the same author has shed light on the emergence of Italian fascism (Franzosi 2010) and on lynching in Georgia (Franzosi, De Fazio and Vicari 2012) using similar techniques.

Along the same lines, the U.S. National Endowment for the Humanities asked the public the question "how can you use open [newspaper archive] data to explore history?" in their Chronicling America Historic American Newspapers Data Challenge (Weinryb Grohsgal 2017), selecting six projects ranging from studies on tracking Biblical quotations (Mullen 2016) and exploration of information in agricultural news (Giroux, Giroux and Galbreath 2016), to discovering patterns of news coverage of the KKK and secession in the U.S. (Palin 2016) that highlight the practical ways in which newspaper archives can be used to touch on a variety of important themes within the humanities.

Work by Nicholson has also explored the methodological possibilities of digital newspaper archives, considering whether media history is on the cusp of a "Digital Turn" (Nicholson 2013), with other large-scale historical studies using computer-assisted techniques providing exemplar studies of related approaches in other historical domains (Dexter et al. 2017; Hughes et al. 2012; Michel et al. 2011).

In this study, we have explored the decades between the 1866 war and the Great War, which were times of increased political polarisation and ethnic awareness in Gorizia, due in part to local immigration and urbanisation issues and in part to a general mood shift in the entire Empire. Additionally, they were also years of crisis and prosperity, recession and growth, tradition and innovation, anxiety for the future and faith in progress. We can see the trace of all that in the news corpus we have digitised, of *L'Eco del Litorale* and *Il Corriere Friulano - Corriere di Gorizia*.

In this data, we can find the individual stories of thousands of people, but also the collective trends of a population heading towards a new chapter in its history. We see new technologies, new ideas, new economic opportunities, new cultural challenges and problems.

Importantly, we get a glimpse in the last years of a world during a period that transformed it beyond recognition. By the time the first shot is fired in Sarajevo, all ingredients of the major conflict are in place: national tensions, powerful technologies, mass media, new ideologies. This small and ancient land, the County of Gorizia, reflected in that period all the general trends of the Empire, perhaps also of Europe.

In this paper we have shown that, in the space of a few decades, the town embraced new ways to communicate, such as the cinema and the telephone, along with modes of transportation, like the car, the airplane, the bicycle and the train. Far from being a backwater of a decaying empire, this was a city with an eye on the future and an interest in new ideas – including political ones. It was, however, also a time in which new tensions emerged along ethnic lines.

The war of course would transform the city and its county into something entirely different. The front lines crossed through the city itself and the urban population was largely relocated. The annexation of the city by Italy was quickly followed by twenty years of fascism, another war, and finally the iron curtain that ran right through the County itself, partly separating the city centre and some of its neighbourhoods.

Today, Gorizia and its younger sister, Nova Gorica, are united by the Schengen agreement and by a common currency, not to mention 1000 years of shared history. It is incredibly fortunate that the collection of newspapers in the Biblioteca Isontina has survived so many threats, to reach us with its unique collection of newspapers. It is very important that we continue to digitise them now, starting from those that are most local and could not be found elsewhere, such as the two Italian newspapers that we started from. We hope that this will be our next project.

New outlets that need to and can be digitised right now include *Il Goriziano, Il Friuli Orientale,* and *Il Gazzettino Popolare*. Fortunately, other Slovenian newspapers from Gorizia can already be found in dlib.si, for example *Primorec*. Adding these new outlets would help us further represent all years and political positions. We could then include those outlets that were printed - at times - in smaller towns of the County, such as Tolmin or Gradisca.

A deeper analysis of these contents would not only shed fresh light on such a key area of southern Central Europe, but would help us refine technical and conceptual tools that would readily transfer to the study of other areas, where rich linguistic and ethnic diversity has shaped centuries of history.

Indeed, since this territory is a unique crossroad, where Slavic, Latin, German and Mediterranean worlds have met for centuries, a digital humanities study of its history could serve as a paradigm for a similar study of all of Europe, towards a shared understanding of our common past.

6 - Notes

1. The first historical document referring to Gorizia is dated 28th April 1001. It is a Latin text mentioning “one village that in the language of the Slavs it is called Goriza” (*unius ville que Sclavorum lingua vocatur Goriza)* (Marušič 2005, 7; Cavazza 2001, 3). In Slovenian, ‘Gorica’ (pron. Goriza) literally means ‘small mountain’.
2. Ger: *Gefürstete Grafschaft Görz und Gradisca*; It: *Principesca Contea di Gorizia e Gradisca*; Sl: *Poknežena grofija Goriška in Gradiščanska*.
3. Ger: *Österreichisches Küstenland*; It: *Litorale Austriaco*; Sl: *Avstrijsko Primorje*.
4. Gorizia was renowned as the ‘Austrian Nice’ since Carl von Czoernig, an Austrian

statistician and historian who chose to spend his retirement age in the Habsburg city, nicknamed it in his book *Das Land Görz und Gradisca* in 1873 (Fabi 1991, 70).

5. This is the only edited edition of the final report of the "Slovene-Italian historical and cultural commission" that worked for the understanding of the Slovene-Italian relations in 1880-1956. The report was printed in Italian, Slovenian and English in 2000. The Commission was established by Italian and Slovenian authorities, and was joined by the most prominent Italian and Slovenian historians on the issue. It met from 1993 to 2000, until a final agreement regarding major topics in the shared history of the two countries was found (Kacin-Wohinz and Troha 2000).

6. For example, *L'Eco del Litorale* was temporarily suspended when it criticised Vienna's joining of the Triple Alliance with Italy in 1882, which had defeated the Papal States in 1871.

7. https://www.abbyy.com/en-gb/finereader12/en/

8. https://www.dlib.si/Help.aspx

9. Snowball stemmer for Italian (http://snowball.tartarus.org/algorithms/italian/stemmer.html) and Slovenian (http://snowball.tartarus.org/archives/snowball-discuss/att-0670/01-slo.proc).

10. We also removed some articles and articled-prepositions (using an elision filter for Italian, contained in the package Lucene) *i.e.* "c", "l", "all", "dall", "dell", "nell", "sull", "coll", "pell", "gl", "agl", "dagl", "degl", "negl", "sugl", "un", "m", "t", "s", "v", "d".

11. From the set of n-grams extracted from the text, the relative probability was estimated by computing a word rank score for each n-gram within each trimester (their position in the lexicon after sorting by frequency). The word rank score $r_w(t)$ was used to estimate the relative frequency of a word $f_w(t)$ by using Zipf's law as

$$f_w(t) \cong r_w(t) = \frac{1}{k_w(t)H_{n(t)}} - \varepsilon(t)$$

$$\varepsilon(t) = \frac{1}{k_\emptyset(t)H_{n(t)}}$$

where $k_w(t)$ is the rank of an n-gram at time interval t, $k_\emptyset(t)$ is the rank of the zero-frequency n-grams with the time interval, $H_{n(t)}$ is the generalised harmonic number for total number of words found within the time interval and $\varepsilon(t)$ is a time dependent correction term used to calibrate time series to the same baseline, *e.g.* for zero frequency n-grams. These word rank scores for each time interval were then assembled into a time series for the given n-gram (Lansdall-Welfare and Cristianini, 2017).

12. For this part of the analysis we removed 'stop words', that is frequent words with just a grammatical meaning, not indicative of a specific topic. These were 619 words for Italian, 446 for Slovenian, obtained from GitHub: https://github.com/6/stopwords-json.

13. *"Delle 900 case che esistono 700 furono gravemente danneggiate e 300 dovranno*

venire demolite". Corriere di Gorizia, 19/04/1895, p. 3.

14. *"La paura della cometa - La fine del mondo! - La comete di Halley ha destato terrore fra le popolazioni rurali slovene e croate della Carniola, del territorio di Trieste e della Dalmazia. Secondo rapporti pervenuti al Governo, la paura è così grande e la convinzione del prossimo finimondo così diffusa che parecchi contadini pensano di vendere i loro beni e di darsi alla pazza gioia che tanto fa lo stesso. Insomma una ripetizione dei terrori del 1000, ma con una rassegnazione più allegra. Ciò posto, il Ministero dell'istruzione mandò un ordinanza ai governatori della Carniola, di Trieste e delle Dalmazia perché provvedano a tranquillare [sic] le popolazioni a mezzo dei maestri e dei parroci, spiegando popolarmente nella scuola a e dal pulpito la teoria delle comete. Un apposito opuscolo si distribuirà ai maestri e si preti"*. "L'Eco del Litorale, 16/04/1910, p. 2.

15. *"Avete mai osservato quelle baracche [...] ambulanti che sono i tram cittadini. [...] Din, din, din: s' avanzano le carrozze col tetto barcollante [...]. I ronzini che vi sono attaccati pare abbiano la maledizione sopra di loro. Procedono a testa bassa come cani battuti, magri, stecchiti, duri, che muovono le gambe a stento. [...] pensavamo nel medesimo tempo alle cose del Municipio così malposte, cosi male condotte. Aumenterà ora il movimento dei forestieri. Dalla stazione della Transalpina a quella delia Meridionale [...] Dove si metteranno questi forestieri? Avrà il coraggio l'amministrazione cittadina di stivarli nelle carrozze di tram [...]? E come va col progetto del nuovo tram? [...] ma purtroppo al Municipio non si grida che per ipotetici o fondati diritti, senza volere riconoscere che vi sono anche dei doveri da compiere."* "Questioni cittadine", L'Eco del Litorale, 18/07/1906, p. 3.

16. *La luce che parla*. "Un importante scoperta", L'Eco del Litorale, 07/10/1880, p.2

17. *"Si potrà d'ora poi in tenere un colloquio fra due persone a oltre 2600 miglia l'una dall'altra parlando 120 parole al minuto, cioè due parole al minuto secondo, rapidità considerevole per le lingue femminili"*. "Un telefono transatlantico", Corriere di Gorizia, 15/09/1883, p. 3.

18. "Il Municipio comunica le seguenti notificazioni:", L'Eco del Litorale, 17/02/1894, p. 3.

19. "Telefono", L'Eco del Litorale, 12/02/1894, p. 2

20. For example: in the Christmas week of 1907, Il Corriere Friulano repeatedly published an ad for the travelling "Grand Elektrik Bioskop of J. Bahmaier" stationed in those days near the city centre, promising at the end: "black evening every saturday at 9pm only for gentlemen over 18". "Grand Elektrik Bioskop", Il Corriere Friulano, 20/12/1907, p. 3.

21. *"È una novità assoluta, che per chi non ne conosce il meccanismo e le leggi fisiche a cui obbedisce, ha del meraviglioso"*. "La

fotografia animata", Corriere di Gorizia, 24/11/1896, p. 1.

22. *"Le vedute di questo cinematografo sono varie; p. e. esso ci mostra […], una bagnante, l'arriva di un treno ferroviario, il movimenta dei passeggieri ecc. Questa della ferrovia è anzi una della veduta più interessanti. Si vede il convoglio in arrivo, poi i conduttori che aprono gli sportelli, la discesa dei paseggieri, a chi si piglia una valigia, chi un cagnolino ecc., tutto molto chiaro e molto bene, tanto che specialmente questo quadro della ferrovia fu calorosamente applaudito".* "Il cinematografo al salone Dreher", Corriere di Gorizia, 10/12/1896, p. 3.

23. Handwritten on original picture of the Rusjan's brothers "*Trapola de Carta*" airplane at: http://www.edvard-rusjan.it/rus005.jpg

24. "*Officina Elettrica*", Il Corriere Friulano, 09/06/1903, p. 2.

25. "*L'illuminazione a luce elettrica dalla città di Gorizia*", Il Corriere Friulano, 15/08/1903, p. 2.

26. *"Ieri girava per la nostra città un automobile che attirava lo sguardo e l'attenzione di tutti. Era montato dei due signori fratelli conti Giulay, i quali avevano fatto su quello il viaggio da Vienna […]. Questo automobile, uno dei più perfetti che esistano, impiega due minuti per ogni chilometro in pianura […]".* "Un automobile modello", Corriere di Gorizia, 06/12/1898, p. 3.

27. *"Martedì alle 16 ebbe luogo una gara tra un automobile a benzina e un velocipedista. Quantunque lo chauffeur corresse a tutta forza, il velocipedista, sino all'Ospitale femminile, aveva un vantaggio di circa 5 metri sull'automobile".* "Un match tra un automobile e un velocipede", Il Corriere di Gorizia, 15/08/1901, p. 3.

28. *"Non vi è che una una voce per dire il malumore contro questa strana indifferenza dei velocipedastri e della calma con cui procede l'autorità nel porvi freno. Tutti i giorni si deplorano fatti gravissimi. L'altro ieri in Piazzutta mancò poco che una ragazza fosse vittima della corsa sfrenata di un velocipedista. Due giorni sono lo stesso toccava in Corso ad una signora e stamane purtroppo una disgrazia avvenne che manco' poco non portasse irreparabili conseguenze. Un amore di bambina d' anni 6 Annita P. si trovava accompagnata dalla serva in P. Grande. Colla vivacità solita dei bambini le sfuggì per recarsi a vedere una baracca. Un velocipedista che veniva in corsa precipitosa da Via Rastello rovesciò la bambina e la ferì in più parti alla faccia ed ai polsi. Fu miracolo se non rimase schiacciata o deturpato per sempre il voltino di quell' angioletto! Si nota altresì che questi prodi si involano invece di fermarsi e almeno mostrare un po' d' interesse alle loro vittime. Il più delle volte è così. Bei tomi! Sono fatti e non ciarle questi e anche troppo eloquenti".* "Lo spavento delle famiglie", Il Corriere di Gorizia, 25/08/1898, p. 2.

29. *"[...] viene generalmente deplorata l'incongruenza di certi deputati della parte slava, i quali [...] si adoperarono con ogni possa per rendere illusoria l'urgente erezione del manicomio per la nostra provincia. [...] l'argomento del fallito manicomio, in cui c'entra una furiosa dissensione fra i deputati delle due nazionalità, pari a quella di Babele, ben persuasi che per ora ogni argomentazione della questione, trita e ritrita, sia post factum affatto superflua".* "Contromisure necessarie", *Corriere di Gorizia*, 07/10/1889, p. 1.

30. *"Esiste la questione sociale, dice il P. Pavisich ed esiste perchè, si voglia o non si voglia, nel mondo attualmente da una parte ravvisiamo infinite dovizie nelle mani di pochi, dall'altra miseria, pauperismo, fame. [...] Colla statistica alla mano dimostra quale sia il depauperamento del popolo operato dall'egoismo dei capitalisti. [...] il P. Pavisich indica nella rivoluzione francese la causa prossima di tale stato intollerabile. [...] Ma ora il popolo si ribella a questo stato di cose, schierandosi in gran massa sotto il rosso vessillo della democrazia e rendendo cosi il conflitto vieppiù aspro, foriero d'una conflagrazione universale".* "La prima conferenza sociale a S. Antonio Nuovo", L'Eco del Litorale, 15/04/1898, p. 1.

31. "[...] *procurare per l'avvenire un'amministrazione giusta, imparziale, non favorevole ad una schiatta al di sopra delle altre*." "Cronaca Politica", L'Eco del Litorale, 03/08/1879, p. 2

32. *Al di sopra dei partiti*. "Note Parlamentari", L'Eco del Litorale, 15/04/1880, p. 1.

33. *"[...] egli si e' prefissata una meta; e che questa meta e' la grandezza dell'impero, e il contentamento dei popoli ond'e' composto, proponendosi di soddisfare le legittime aspirazioni dei popoli medesimi".* "Note Parlamentari", L'Eco del Litorale, 15/04/1880, p. 1.

34. *"Se dunque i tedeschi dell'Austria posseggono un Schulverein, se gli slavi hanno la loro società dei santi Cirillo e Metodio, ambedue dette società sorte per proteggere non solo, ma propagare (ho detto propagare) le due lingue, perché noi Italiani appartenenti allo stato medesimo ed aventi quindi gli stessi diritti, non possiamo avere una società "Pro Patria" la quale protegga e coltivi il dolce idioma nostro?".* "Per l'Inaugurazione del Gruppo Locale della Società Pro Patria", Corriere di Gorizia, 20/09/1887, p. 1.

35. http://www.catasti.archiviodistatotrieste.it/Divenire/collezione.htm?idColl=10649282

7 - References

Agostinetti, N. 1981. L'Attività dei Cattolici Isontini nel Primo Ventennio del Novecento. In *I Cattolici Isontini nel XX Secolo - I - Dalla Fine dell'800 al 1918*, Gorizia, Le Casse Rurali e Artigiane della Contea di Gorizia.

Allen, R.B., I. Waldstein, and W. Zhu. 2008. Automated Processing of Digitized Historical

Newspapers: Identification of Segments and Genres. In *International Conference on Asian Digital Libraries* (pp. 379-386). Springer, Berlin, Heidelberg.

Cavazza, S. 2001. *Gorizia e il Territorio: ConsIderazioni intorno al Millenario Goriziano*, Il Territorio (Monfalcone), 16.

Cvrček, T. 2013. Wages, Prices, and Living Standards in the Habsburg Empire, 1827–1910, *Journal of Economic History*, 73, 1 (March), 1-37.

De Grassi, M. 1982. Catalogo dei periodici stampati o editi nella Contea di Gorizia e Gradisca conservati nelle biblioteche pubbliche isontine (1774-1918), *Studi Goriziani*, 55-56, pp. 51-104.

De Simone, G. 1996. Catalogo dei periodici posseduti in microfilm dalla Biblioteca statale isontina, *Studi Goriziani*, v. LXXXIV, luglio-dicembre, pp 131 - 144.

Dexter, J.P., T. Katz, N. Tripuraneni, T. Dasgupta, A. Kannan, J.A. Brofos, J.A. Bonilla Lopez, L.A. Schroeder, A. Casarez, M. Rabinovich, A.H. Lushkov. 2017. "Quantitative criticism of literary relationships." *Proceedings of the National Academy of Sciences* 114, no. 16: E3195-E3204.

dLib.si 2017. Digital Library of Slovenia, http://dlib.si

Fabi, L. 1991. *Storia di Gorizia*, Padova Il Poligrafo.

Feresin, V. 2007-2008. La Stampa a Gorizia – Fra Settecento e Ottocento, *Isonzo Soča*, n° 75-76, 14-21.

Ferrari, L. 2002. Gorizia Ottocentesca, fallimento del progetto della Nizza Austriaca. In *Storia d'Italia. Le regioni dall'Unità ad oggi. Il Friuli Venezia Giulia - vol. I*, edited by R. Finzi, C. Magris, and G. Miccoli, 313-375. Torino, Einaudi.

Filipič, I. 2010. Stepišnik in Sveta Brata Ciril. In *Metod, Bogoslovni vestnik, n°70*.

Fogel, R.W., and G.R. Elton. 1983. Which road to the past?: two views of history. Yale University Press.

Franzosi, R. 2010. Sociology, narrative, and the quality versus quantity debate (Goethe versus Newton): Can computer-assisted story grammars help us understand the rise of Italian fascism (1919–1922)?. *Theory and society*, 39(6), pp.593-629.

Franzosi, R. 2017. A third road to the past? Historical scholarship in the age of big data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50(4), pp.227-244.

Franzosi, R., G. De Fazio, and S. Vicari. 2012. Ways of measuring agency: an application of quantitative narrative analysis to lynchings in Georgia (1875–1930). *Sociological Methodology*, 42(1), pp.1-42.

Garimberti, C. 1877. *Diario Storico del Viaggio di Francesco Giuseppe I a Trieste, Gorizia, Venezia,*

in Istria, in Dalmazia ed a Fiume , 1875, Zara, Vitaliani & Janković.

Giroux, A., N. Giroux, and M. Galbreath. 2016. Historical Agricultural News. [ONLINE] Available at: http://ag-news.net. [Accessed 7 February 2018].

Gorian, R. 2010. *Gazzetta Goriziana Editoria e informazione a Gorizia nel Settecento*, Trieste, Deputazione di storia patria per la Venezia Giulia.

Horel, C. 2015. Austria-Hungary 1867-1914. In, *Political Censorship of the Visual Arts in Nineteenth-Century Europe*, by R.J. Goldstein, A.M. Nedd. Basingstoke, Palgrave Macmillan.

Hughes, J.M., N.J. Foti, D.C. Krakauer, and D.N. Rockmore. 2012. Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, 109(20), pp.7682-7686.

Impact 2008. IMPACT project - Improving access to text, information available at http://www.impact-project.eu/

Inoue, N. 2008. *Le industrie goriziane e l'istituto per la promozione delle industrie di Gorizia 1903-1914,* Doctoral Thesis, University of Trieste.

Jerele, I., T. Erjavec, D. Pokorn, and A. Kavčič-Čolić. 2011. Optical character recognition of historical texts: end-user focused research for Slovenian books and newspapers from the 18th and 19th century. *SEEDI Conference: Proceeding* 2011. - https://www.digitisation.eu/download/website-files/lexica/NCD21117.pdf

Kacin-Wohinz, M., and N. Troha. 2000. *Slovensko-italijanski odnosi 1880-1956. Poročilo slovensko-italijanske zgodovinsko-kulturne komisije / Rapporti italo-sloveni 1880-1956. Relazione della commissione storico-culturale italo-slovena / Slovene-Italian relations 1880-1956. Report of the Slovenian-Italian historical and cultural commission*, Ljubljana, Nova revija.

Kalc, A. 2013. Vidiki razvoja prebivalstva Goriške-Gradiške v 19. stoletju in do prve svetovne vojne / Some aspects of the demographic development in Goriška-Gradiška from early 19th century to WWI, *Acta Histriae,* 4(21).

Lansdall-Welfare, T., and N. Cristianini. 2017. History Playground: A Tool for Discovering Temporal Trends in Massive Textual Corpora. *Manuscript submitted for publication.*

Lansdall-Welfare, T. et al. 2017. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences* 114(4) E457-E465, doi:10.1073/pnas.1606380114.

Marušič, B. 2005. *Pregled politične zgodovine Slovencev na Goriškem – 1848-1899*, Nova Gorica, Goriški Muzej.

Medeot, C. 1981. Panorama Politico. In *I Cattolici Isontini nel XX Secolo - I - Dalla Fine dell'800 al 1918,* Gorizia, Le Casse Rurali e Artigiane della Contea di Gorizia.

Michel, J.B. et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182.

Mlakar, L., and A. Turel. 2010. *Storia di Gorizia*, Pordenone, Biblioteca dell'immagine.

Moretti, F. 2013. *Distant Reading*, London,Verso.

Mosse, G. 1974. *The nationalization of the masses. Political symbolism and mass movements in Germany from the Napoleonic wars through the Third Reich*, New York, H. Fertig.

Mullen, L. 2016. America's Public Bible: Biblical Quotations in U.S. Newspapers, website, code, and datasets. [ONLINE] Available at: http://americaspublicbible.org. [Accessed 7 February 2018].

Nicholson, B. 2012. Counting culture; or, how to read Victorian newspapers from a distance. *Journal of Victorian Culture*, *17*(2), 238-246.

Nicholson, B. 2013. The Digital Turn: Exploring the methodological possibilities of digital newspaper archives. *Media History*, 19 (1). pp. 59-73. ISSN 1368-8804 DOI 10.1080/13688804.2012.752963

Palin, R. 2016. Digital History Projects in the AP U.S. History Classroom. [ONLINE] Available at: https://apush.omeka.net. [Accessed 7 February 2018].

Patat, L. 2003. *Fra Austria e Italia - Cormòns e l'Isontino a cavallo di due secoli*, Udine, IFSML.

Powers, D.M. 1998. Applications and explanations of Zipf's law. *Proceedings of the joint conferences on new methods in language processing and computational natural language learning* (pp. 151-160). Association for Computational Linguistics.

Redivo, D. 2005. *Le trincee della Nazione: cultura e politica della Lega Nazionale (1891-2004)*, Trieste, Edizioni degli Ignoranti Saggi.

Scandolara, S. 2001. *Nostro cine quotidiano: le Gorizie al cinema*, Gorizia, Kinoatelje.

Weinryb Grohsgal, L. 2017. Using Big Data to Ask Big Questions: A Digital Humanities Challenge in Historic Newspapers, paper presented to Digital Humanities 2017, Montréal, 8-11 August, < https://dh2017.adho.org/abstracts/374/374.pdf >